Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 6

A Theory of Reciprocity

Armin Falk and Urs Fischbacher

July 2000

A Theory of Reciprocity

by

Armin Falk and Urs Fischbacher*
University of Zurich
Institute for Empirical Economic Research Bluemlisalpstrasse 10, CH-8006 Zürich
falk@iew.unizh.ch, fiba@iew.unizh.ch

First version: March 1998
This version: July 2000
Keywords: Reciprocity, Fairness, Cooperation, Competition, Game Theory.
JEL: C7, C91, C92, D64, H41

Abstract

This paper presents a formal theory of reciprocity. Reciprocity means that people reward kind actions and punish unkind ones. The theory takes into account that people evaluate the kindness of an action not only by its consequences but also by the intention underlying this action. The theory explains the relevant stylized facts of a wide range of experimental games. Among them are the ultimatum game, the gift-exchange game, a reduced best-shot game, the dictator game, the prisoner's dilemma, public goods games, and the investment game. Further, the theory explains why subjects behave differently in treatments where they experience the actions of real persons compared to treatments where they face 'actions' caused by a random device. Finally, the theory explains why in bilateral interactions outcomes tend to be 'fair' whereas in competitive markets even extremely unfair distributions may arise.

*Financial support by the Swiss National Science Foundation (Project 12-43590.95) and by the MacArthur Foundation (Network on Economic Environments and the Evolution of Individual Preferences and Social Norms) is gratefully acknowledged. We would like to thank Sam Bowles, Martin Brown, Simon Gächter, Herbert Gintis, Lorenz Götte, Matthew Rabin, Armin Schmutzler and the participants of MacArthur Foundation meeting in Chicago 1998.

Kindness is the parent of kindness.
(Adam Smith)

1 Introduction

A large body of evidence indicates that reciprocity is a powerful determinant of human behavior. Experiments and questionnaire studies performed by psychologists and economists as well as an impressive literature in sociology, ethnology and anthropology emphasize the omnipresence of reciprocal behavior. The sociologist Gouldner (1960), for example, notes that the norm of reciprocity is "no less universal and important an element of culture than the incest taboo..." (p. 171).

The importance of reciprocity for economics has been pointed out by many scholars. Kahneman, Knetsch, and Thaler (1986), e.g., show that fairness norms prevail in a variety of business contexts. In the field of labor economics, questionnaire studies with owners and managers of firms suggest that a possible source for rigid wages is that the employers are unwilling to cut wages (Blinder and Choi (1990), Bewley (1995), Agell and Lundborg (1995), Campbell and Kamlani (1997) and Agell (1999). According to these studies a major reason for firms' refusal to cut wages is the fear that pay cuts will adversely affect work morale. Thus, concerns for reciprocity may play a key role in the explanation of downwardly rigid wages. Economic consequences of reciprocity have also been shown in many other areas, such as tax compliance (Smith (1992)), organization theory (Steers and Porter (1991), Mowday (1991)), contributions to public goods (Sugden (1984)), contract enforcement (Fehr and Falk (1999)), gift-giving (Ruffle (1995)), and strike breaking (Francis (1985)).

The essence of reciprocity is very nicely captured in a quote from *The Edda*, a medieval collection of Icelandic epic poems: "A man ought to be friend to his friend and repay gift with gift. People should meet smiles with smiles and lies with treachery." The quote includes *positive reciprocity* which is the reward of a kind treatment and *negative reciprocity* which means the punishment of an unkind treatment. Importantly, reciprocity means a behavior that cannot be justified in terms of selfish and purely outcome oriented preferences. To avoid terminological confusion let us, therefore, clarify that reciprocity sharply distinguishes from 'reciprocal altruism' (Trivers (1971)). A reciprocal altruist is only willing to reciprocate if there are future rewards arising from reciprocal actions.[1]

Experimentalists have found evidence for both negative as well as positive reci-

[1] In the parlance of game theory this kind of reciprocal action may be supported as an equilibrium strategy in infinitely repeated games (folk theorems) or in finitely repeated games with incomplete information (see Kreps, Milgrom, Roberts, and Wilson (1982)).

procity: Negative reciprocity is a very robust observation, e.g., in the so-called ultimatum game.[2] Starting with the work of Güth, Schmittberger, and Schwarze (1982) numerous studies have shown that people - contrary to their material self-interest - reject low offers in order to punish the unkindness of proposers. Positive reciprocity has been found, e.g., in the investment game studied by Berg, Dickhaut and McCabe (1995) and in the gift-exchange game by Fehr, Kirchsteiger, and Riedl (1993). In the latter, the authors analyze the impact of reciprocity in an experimental labor market. They report a significantly positive relationship between wages paid by firms and the effort level provided by workers. Besides this obvious evidence for reciprocity, there is also evidence which *seems to be incompatible with reciprocal preferences*: Many market experiments, e.g., impressively support the outcome predicted by the standard economic theory which assumes completely selfish preferences. In this paper we present a theory of reciprocity which explains the relevant stylized facts of a wide variety of experimental games. This includes both, "fair" outcomes in bargaining games as well as "unfair" outcomes in markets.

According to our theory, a reciprocal action is modeled as the behavioral response to an action that is perceived as either kind or unkind. The more an action is considered as kind or unkind, the more it will be rewarded or punished, respectively. The crucial question to ask is: *How do people evaluate whether an experienced action is kind or unkind?* Two aspects are essential here, namely (i) the *outcome* or the consequences of an action and (ii) the underlying motivation, i.e., the *intentions* involved. It is not only the consequence that determines the kindness of an action. Rather, "people determine their dispositions toward others according to motives attributed to these others, not solely according to actions taken" (Rabin (1998), p. 22). As an example take the criminal law which distinguishes quite carefully between criminal activities that were intended and those that were not. A particular crime is considered less criminal and will be punished less if it was committed negligently and not with criminal intent (compare, e.g., the concepts of manslaughter and second-degree manslaughter). As another example take again pay cuts of a firm: As Bewley (1995) shows in his questionnaire study, workers interpret pay cuts as an insult which adversely affect work morale. This, however, may hold to a much lesser degree if workers know that their firm is "forced" to lower wages, e.g., to avoid bankruptcy.

Several other attempts to model other-regarding preferences have been made. Most recently, Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) have suggested theories which emphasize inequity aversion as a motivational determinant. These models are purely outcome-oriented in the sense that they do not account

[2] All games mentioned in this section are explicitly described in the applications section.

for the role of intentions. Consequently, they ignore an important aspect of the psychology of reciprocity. Moreover, the models fail to predict important experimental results (see our discussion in Section 2).

A different fairness model has been developed by RABIN (1993).[3] His model explicitly accounts for the role of intentions. However, Rabin's strategic form theory is not meant to predict the data of sequential games.[4] In the sequential prisoner's dilemma, e.g., Rabin's theory predicts unconditional cooperation and rules out conditional cooperation. In the ultimatum game his model is compatible with offers above 50 percent of the total pie. Both predictions are refuted by the data.

In our theory, we implement an equity based reference standard for the evaluation of an outcome's kindness. However, we also take into consideration that in judging the kindness of an action, people care for fair intentions. The theory yields testable predictions. We show that it explains the stylized facts of a wide variety of experimental games. We derive all predictions with a *single* utility function. Furthermore, we use the *same* parameter constellation in all games. This is remarkable since we do not only explain outcomes of games where people behave in a "fair" and other-regarding way. Rather, we also show and explain why the same subjects act as if they were solely driven by self-interest. In particular, our theory predicts the stylized facts of the following games, the ultimatum game, the dictator game, the gift-exchange game, the prisoner's dilemma, a reduced best-shot game, the investment game, public good games and competitive market games.

The remainder of this paper is organized as follows: In Section 2 we contrast the two distinct concepts of reciprocity and inequity aversion. Section 3 introduces the formal theory. Section 4 provides applications of the theory in the light of the experimental evidence. Section 5 summarizes.

2 Why inequity aversion reaches too short

Recently two models of other-regarding preferences have reached considerable attention: the inequity aversion models by BOLTON AND OCKENFELS (2000) and FEHR AND SCHMIDT (1999). In their models it is the dislike of inequitable distributions which triggers behavioral responses. *Inequity aversion* sharply contrasts from *reciprocity* for at least two reasons. *First*, inequity aversion is a purely consequentialistic concept, i.e., intentions or motives play no role. Reciprocity on the other hand emphasizes the importance of intentions. *Second*, an inequity averse person punishes or

[3] Other models that take account of fairness preferences are LEVINE (1998) and DUFWENBERG AND KIRCHSTEIGER (1998).

[4] Commenting on his own model Rabin notes that "[e]xtending the model to sequential move games is also essential for applied research." (RABIN (1993), p. 1296).

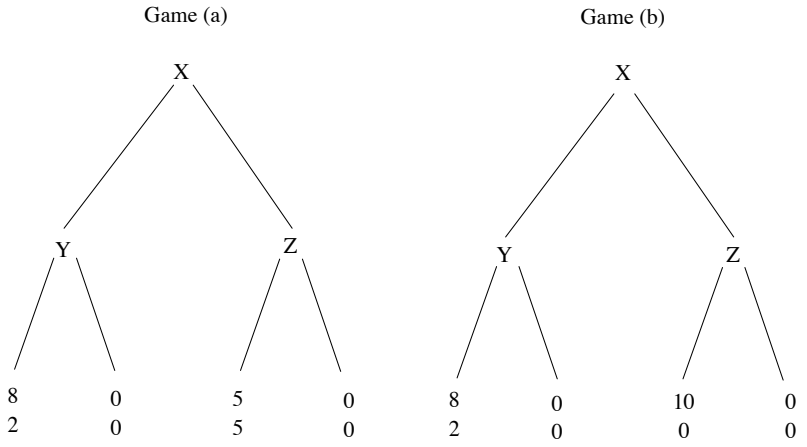

Figure 1: Two games that show the importance of intentions.

rewards another person if and only if this reduces the inequity between the person and his opponent(s). Reciprocity on the other hand means the rewarding or punishment of kindness or unkindness. Therefore reciprocity predicts punishments or rewards even in situations where inequity cannot be reduced. Ultimately it is an empirical question whether inequity aversion or reciprocity is the more powerful concept. In this section we therefore sketch some of the experimental evidence which aims at answering this question.

2.1 Intentions matter

...the intention, not the matter makes the benefit.
(Seneca)

There is an ongoing debate whether intentions are behaviorally relevant. The consequentialistic perspective taken by the inequity aversion approach claims that fairness driven punishment or reward can fully be accounted for without taking intentions into consideration. According to Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) the distributive consequences of an action are sufficient to trigger behavior and no consideration of intentions is needed. Our reciprocity model on the other hand incorporates people's concerns for intentions. To illustrate the role of intentions consider the following two games:

Both games in Figure 1 are reduced ultimatum games. In *game (a)* the first mover is asked to divide 10 dollars between himself and a responder. He has only two choices, either to offer 2 dollars to the responder and to keep 8 dollars (8/2-offer) or to offer an equal split. The responder can either accept or reject the offer. A rejection leads to zero incomes for both players. Consider now *game (b)*. Again, the

proposer's task is to divide 10 dollars. The two possible offers he can choose from are the 8/2-offer, just as in *game (a)* and a 10/0-offer which leaves nothing to the responder.

In our discussion we concentrate on the rejection behavior of the 8/2 offer. The consequentialistic perspective advocated by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) implies that the rejection rates of the 8/2-offer is the *same* in both games. The reason is that these models disregard that identical offers may be perceived as more or less fair, depending on the intentions involved. This view is clearly rejected by the experimental evidence. Even though the *consequences are exactly the same*, responders reject the 8/2-offer significantly more often in *game (a)* than in *game (b)* (compare Falk, Fehr, and Fischbacher (1999)). While the 8/2-offer in *game (a)* is rejected in 44.4 percent of the cases, the same offer is rejected in only 8.9 percent of the cases in *game (b)*. The reason for the different rejection pattern is that responders care about *why* the proposer chose the 8/2-offer, i.e., they care about the proposer's intentions. In *game (a)* it is obvious that the proposer did not want to reach a fair allocation between himself and the responder. He *could have* chosen an equal split but he did not. Offering the 8/2-offer in *game (b)*, however, does not signal bad intentions. After all it is the most 'fair' offer the proposer could offer at all.

The difference in rejection rates between *games (a)* and *(b)* clearly demonstrates the importance of intentions.[5] However, the fact that the rejection rate of the 8/2-offer is greater than zero even if this offer is actually a 'fair', indicates that intentions alone cannot be the whole story either.[6] This view is supported by the experiments by Blount (1995) and Charness (1996) who report that reciprocity is (i) much weaker but (ii) not zero in a condition where intentions are absent compared to a condition where intentions are present. In our model we therefore incorporate *both*, the concern for the outcome per se and for the underlying motivation, i.e., intentions.

2.2 People punish even if they cannot reduce inequity

According to the inequity approach, a person will punish another person if and only if this reduces the inequity between the person and his opponent(s). Reciprocity on the other hand dictates to punish in order to reciprocate an unkind act. The aim of

[5] Experimental evidence for the importance of intentions is found in Bolle and Kritikos (1998), Falk, Fehr, and Fischbacher (1999), Charness (1996), Blount (1995), Greensberg and Frisch (1972), and Goranson and Berkowitz (1966). In the latter study, e.g., the authors report an experiment where people could reciprocate earlier help. Reciprocation was significantly higher when the prior help was provided voluntarily compared to involuntary help.

[6] This questions the approach taken by Dufwenberg and Kirchsteiger (1999) or Rabin (1993) which claims that reciprocity is exclusively intentions driven.

the reciprocating subject is not to reduce inequity but to lower the opponent's payoff. Reciprocity driven punishments are therefore not restricted to situations where inequity can be reduced. Instead it occurs whenever a person is treated unkindly and is given a chance to pay back.

Experimental evidence suggests that many subjects in fact punish others even if punishment does not reduce inequity. Falk, Fehr and Fischbacher (2000) present three experiments which address this question in great detail. As it turns out, a substantial amount of punishment occurs even in situations where inequity cannot be reduced. One of these games is a three person prisoner's dilemma with a subsequent punishment opportunity. In the punishment stage, subjects could – at a cost – deduct points from each of the other two players. In a situation where punishing even *increased* inequity, 46.8 percent of the subjects who had cooperated in the prisoner's dilemma nevertheless punished. The conclusion from the three experiments is that much of the observed punishments are actually triggered as a response to unkindness, not as an attempt to reduce inequity.

The reported evidence suggests that reciprocity is a more powerful concept than inequity aversion. This is so because it better captures the psychological motivations. As a consequence it also yields better predictions.

3 A theory of reciprocity

Our theory formalizes the basic structure of reciprocity which consists of a kind (or unkind) treatment by another person (represented by the *kindness term* φ) and a behavioral reaction to that treatment (represented by the *reciprocation term* σ). Our procedure is to transform a standard game into a psychological game, the so-called "reciprocity game". In this new game the players' utility depends not only on the payoffs of the original standard game but also on the kindness and the reciprocation term. In the following, we derive both terms.

Consider a two-player extensive form game with a finite number of stages and with complete and perfect information. (For notational simplicity we develop the theory for the two-player case. The extension to n-person games is given in Appendix 2.) Let i be a player in the game. N_i denotes the set of nodes where player i has the move with n being a node of this player. Let A_n be the set of actions in node n. Let F be the set of end nodes of the game. The payoff function for player i is given by $\pi_i : F \to \mathbb{R}$.

Let $P(A_n)$ be the set of probability distributions over the set of actions in node n. Then $S_i = \prod_{n \in N_i} P(A_n)$ is player i's behavior strategy space. Thus, a player's behavior strategy puts a probability distribution on each of the player's decision

nodes. Let player j be the other player with behavior strategy space S_j.[7] For $s_i \in S_i$ and $s_j \in S_j$ we define $\pi_i(s_i, s_j)$ and $\pi_j(s_i, s_j)$ as the players' expected payoffs, given strategies s_i and s_j. Furthermore, we define $\pi_i(n, s_i, s_j)$ as the expected payoff conditional on node n: It is the expected payoff of player i in the subgame starting from node n, given that the strategies s_i and s_j are played and given it is known that player i is at node n.[8]

Let s_i' denote the **first order belief** of player i. It captures i's belief about the behavior strategy $s_j \in S_j$ which player j will choose. Similarly, the **second order belief** s_i'' of player i is defined as player i's belief about player j's belief about which behavior strategy player i will choose. In other words, s_i'' is i's belief about s_j'. Like Rabin (1993), we assume that s_i' is an element of S_j and s_i'' is an element of S_i. A set of beliefs is said to be **consistent**, if $s_i = s_j' = s_i''$ holds for $i \neq j$.

3.1 The kindness term φ

The kindness term φ is the central element of our theory. It measures how kind a person perceives the action by another player. As outlined in the previous section the perceived kindness depends on the outcome and the intention. The outcome is measured with the **outcome term** Δ where $\Delta > 0$ expresses an advantageous outcome and $\Delta < 0$ expresses a disadvantageous outcome. In order to determine the overall kindness, Δ is multiplied with the **intention factor** ϑ. This factor is a number between zero and one, where $\vartheta = 1$ captures a situation where Δ is induced fully intentionally and $\vartheta < 1$ expresses a situation where less or no intentions are involved. The kindness term φ is simply the product of Δ and ϑ. All terms are derived in the following.

First, we define the **outcome term**:

$$\Delta(n) := \pi_i(n, s_i'', s_i') - \pi_j(n, s_i'', s_i') \tag{1}$$

To interpret this expression, let us fix the intention factor ϑ. For a given ϑ, the outcome term Δ captures the kindness of player j: The kindness of player j in node n is, ceteris paribus, higher, the more she offers to player i. This is expressed in the term $\pi_i(n, s_i'', s_i')$. From player j's perspective, player j is offering $\pi_i(n, s_j', s_j)$ to player i. This is the payoff player i is expected to get, if player j chooses to play s_j and expects player i to choose s_j'. Put differently, this is - from the perspective of player j - what player j offers to player i, given her expectation about s_i. Player i's belief about this offer is $\pi_i(n, s_i'', s_i')$.

[7] Throughout the paper we will use the male form for player i (and for first movers). For player j (and second movers) we will use the female form.

[8] Because we are working with behavior strategies, the strategies s_i and s_j induce a strategy in the subgame.

Since ϑ is positive, the sign of the kindness term, i.e., whether an action is considered as kind or unkind, is determined by the sign of Δ. In order to determine the sign of Δ, player i needs to compare the offer $\pi_i(n, s_i'', s_i')$ with a *reference standard*. In principle, there are many possible reference standards against which subjects can compare a particular outcome. From many experiments we know, however, that an *equitable* share of payoffs is a salient and commonly held standard. [9] The expression $\pi_j(n, s_i'', s_i')$ serves that purpose. It is player i's belief about what payoff player j wants to keep for herself. Taken together, if $\pi_i(n, s_i'', s_i') > \pi_j(n, s_i'', s_i')$ holds, player i thinks that player j wants him to get more out of the exchange than player j wants for herself, i.e., player i believes that player j is acting kindly. If, on the other hand, $\pi_i(n, s_i'', s_i') < \pi_j(n, s_i'', s_i')$ holds, player i believes that player j claims more for herself than she is willing to leave for player i. In this case, player i perceives player j as being unkind.

Note that we talk a bit loosely about the kindness of an *action*. The way we model kindness comprises both the kindness of actually occurred actions as well as anticipated future actions. The actions which already occurred are represented in node n because being in node n results from player j's actually chosen actions. The anticipated actions of player j are simply captured by s_i'.

The outcome term $\Delta(n)$ captures the kindness or unkindness of player j towards player i for a given ϑ. In judging player j's kindness, however, player i is well aware of the fact that player j might not have caused a particular outcome intentionally. The Δ-term does not differentiate between situations where player j had a reasonable alternative and those where he had none. In order to account for this difference, $\Delta(n)$ is multiplied with ϑ which depends on player j's set of alternatives. Let us state precisely what we mean by alternative payoff combinations. Let S_j^p be the set of pure strategies of player j. For given strategies and beliefs we define:

$$\Pi_i := \left\{ (\pi_i(s_i'', s_j^p), \pi_j(s_i'', s_j^p)) \mid s_j^p \in S_j^p \right\} \tag{2}$$

Π_i is a set of payoff combinations π_i and π_j. These are payoffs player j can induce by choosing a pure strategy s_j^p given her expectation about player i's strategy. Since Π_i is determined from player i's perspective, player i takes into account his belief about which strategy player j believes he will choose, namely s_i''. In short, Π_i is player i's belief about all payoff combinations player j considers as her opportunity set, i.e., given s_i'' it contains all options available to player j.

[9] The idea that equity considerations are important in the context of judging kindness was first developed in the so-called *equity theory*. Beginning in the late sixties social psychologists developed *equity theory* as a special form of *social exchange theory*. Compare, e.g., Adams (1965) and Walster and walster (1978). See also Loewenstein, Thompson and Bazerman (1989) and Selten (1978).

Whether player i experiences a particular outcome as chosen intentionally depends on the options available to player j. The question to ask is: How intentional is a particular payoff distribution (π_i^0, π_j^0) induced by player j - given that player j had an alternative payoff distribution (π_i, π_j) he could have chosen? The answer to this question is given in function Ω. In this function, the value of Ω expresses how intentional player j's choice of (π_i^0, π_j^0) was, given his alternatives (π_i, π_j). If the choice was fully intentional Ω equals 1, if the choice is considered as not fully intentional, however, Ω is smaller than one. The Ω-function is defined as follows:

$$\Omega(\pi_i, \pi_j, \pi_i^0, \pi_j^0) := \begin{cases} 1 & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i < \pi_i^0 \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i \geq \pi_i^0 \\ 1 & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i \leq \pi_j \\ \max\left(1 - \frac{\pi_i - \pi_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i\right) & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i > \pi_j \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \pi_i \leq \pi_i^0 \end{cases} \quad (3)$$

where ε_i is an individual parameter with $0 \leq \varepsilon_i \leq 1$.[10]

The first two rows capture situations where player j has treated player i in a kind way ($\pi_i^0 \geq \pi_j^0$). In these situations the value of Ω depends on whether player j could reduce player i's payoff (π_i compared to π_i^0) or not. Assume, as shown in the first row, player j *has* the alternative to lower player i's payoff (i.e., $\pi_i < \pi_i^0$). Then player i considers the kind action as fully intentional. After all, player j could have treated player i worse but chose not to do so. If, however, player j's only alternative is to increase player i's payoff (i.e., $\pi_i \geq \pi_i^0$) then player j has *no alternative to be less kind*. Consequently, his kindness is considered as less intentional. After all, it is not player j's merit that he was 'kind'. In this situation Ω equals ε_i.

The other three rows represent instances where player j puts player i in a disadvantageous situation, i.e., where $\pi_i^0 < \pi_j^0$ holds. Whether player i in fact perceives this as an intentionally unkind action depends again on player j's available alternatives. If player j *has* the alternative to improve player i's payoff without putting himself in a disadvantageous situation ($\pi_i > \pi_i^0$ and $\pi_i \leq \pi_j$), his unkindness is fully intentional. Therefore, Ω is equal to 1. Now suppose that there *is* an alternative to improve player i's payoff which at the same time leads to a disadvantageous situation for player j. The more this alternative is disadvantageous for player j, the less reasonable it is considered. After all it would imply an *unreasonable sacrifice* by player j. As a consequence, the choice of π_i^0 is not considered as fully intentionally unkind and Ω is equal to $\max\left(1 - \frac{\pi_i - \pi_j}{\pi_j^0 - \pi_i^0}, \varepsilon_i\right) \leq 1$. The expression $1 - \frac{\pi_i - \pi_j}{\pi_j^0 - \pi_i^0}$ measures 'how far' player j must switch to the disadvantageous side if she wants to improve player i's

[10] This parameter is called the pure outcome concern parameter. It is interpreted below.

payoff - related to the reference situation (π_i^0, π_j^0). If, e.g., player j must only switch a little bit to the disadvantageous side (the numerator is small) the alternative action will *ceteris paribus* be considered as rather reasonable. If, however, the numerator is large (in particular, if the numerator is larger than the denominator) Ω is equal to ε_i. Finally, if player j's only alternative is to choose an even lower payoff for player i, i.e., $\pi_i \leq \pi_i^0$, player i cannot infer that player j wanted to treat him in an unkind fashion. Consequently, the action was unintentionally 'unkind' yielding $\Omega = \varepsilon_i$.

As the reference distribution (π_i^0, π_j^0) we use the payoffs that determine the outcome term $\Delta(n)$, namely $\pi_i(n, s_i'', s_i')$ and $\pi_j(n, s_i'', s_i')$. Thus, we define the intention factor:

$$\vartheta(n) = \max\left\{\Omega(\pi_i, \pi_j, \pi_i(n, s_i'', s_i'), \pi_j(n, s_i'', s_i')) \mid (\pi_i, \pi_j) \in \Pi_i\right\} \tag{4}$$

The maximum-operator guarantees that a particular action is considered as intentional if there is *any* 'true' alternative. We now come to the definition of the kindness term.

Definition 1 Let strategies and beliefs be given. We define the **kindness term** $\varphi(n)$ in a node $n \in N_i$ as:

$$\varphi(n) = \vartheta(n)\Delta(n) \tag{5}$$

Interpretation: According to Definition 1 and equations (3) to (5), the kindness term is equal to $\Delta(n)$ with the exceptions mentioned in our discussion of Ω. In these exceptions the kindness or unkindness, respectively, is not fully intentional in the sense that player j had no (reasonable) alternative(s). Therefore, in these cases, the kindness term is *reduced* by a factor between ε_i and 1. Let us illustrate the logic of the kindness term with an example: Imagine player j has three alternative payoff shares (π_i,π_j) he can cause namely (8,2), (7,3) and (6,4). In all these situations player i is favored, i.e., the outcome term Δ is positive. Suppose player j chooses either (8,2) or (7,3). From player i's perspective both choices are fully intentional (with $\Omega = \vartheta = 1$) for player j could have chosen a worse alternative, namely (6,4). Of course, the Δ-term (and therefore the kindness) is higher if j chooses (8,2) instead of (7,3). Now suppose player j chooses the alternative (6,4). Still player i is favored. However, the kindness term is smaller than in the situation where (8,2) or (7,3) is chosen. Moreover, player i knows that player j had no alternative to treat him worse. Thus he cannot infer any positive intentions, i.e., $\Omega = \vartheta = \varepsilon_i$, resulting in an even lower kindness term. Note that in the special case where player j has no alternative at all (i.e., $|\Pi_i| = 1$), $\vartheta(n)$ equals ε_i. In most games, however, the players have a rather unlimited strategy implying a $\vartheta(n)$ of 1.

The individual parameter ε_i is called the **pure outcome concern parameter**. It measures a player's *pure* concern for an equitable *outcome*: If, e.g., ε_i is equal to

zero, player i considers a particular outcome only as kind or unkind if it was caused intentionally, i.e., if the other player had an alternative to act differently. If, on the other hand, $\varepsilon_i = 1$, player i cares *only* about an equitable outcome, i.e., he does not pay any attention to the other player's alternatives and intentions, respectively. In this case Ω (and therefore ϑ) is always equal to 1. Thus, saying that a person's ε_i is equal to 1 is equivalent to saying that this person is purely outcome oriented as it is suggested by the models of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). Insofar, their models can be viewed as a special case of our theory.

3.2 The reciprocation term σ

The second ingredient of our theory concerns the formalization of reciprocation. Let us fix an end node f that follows node n. Then we denote by $\nu(n, f)$ the unique node that directly follows node n on the path that leads from n to f.

Definition 2 Let strategies and beliefs be given as above. Let i and j be the two players and n and f be defined as above. Then we define

$$\sigma(n, f) := \pi_j(\nu(n, f), s_i'', s_i') - \pi_j(n, s_i'', s_i') \tag{6}$$

as the **reciprocation term** of player i in node n.

The *reciprocation term* expresses the response to the experienced kindness, i.e., it measures how much player i alters the payoff of player j with his move in node n. Given player i's belief about player j's expectations about her payoff in node n (i.e., given $\pi_j(n, s_i'', s_i')$), player i can - in node n - choose an action. The reciprocal impact of this action is represented as the *alteration* of player j's payoff from $\pi_j(n, s_i'', s_i')$ to $\pi_j(\nu(n, f), s_i'', s_i')$ (always from player i's perspective). For a given $\pi_j(n, s_i'', s_i')$, player i can, thus, choose to either reward or to punish player j. A rewarding action implies a positive, whereas a punishment implies a negative *reciprocation term*.

Notation: Let n_1 and n_2 be nodes. If node n_2 follows node n_1 (directly or indirectly), we denote this by $n_1 \rightarrow n_2$.

3.3 The utility function

Having defined the kindness and reciprocation term we can now derive the players' utility of the transformed "reciprocity game":

Definition 3 Let player i and j be the two players of the game. Let f be an end node of the game. We define the utility in the transformed reciprocity game as:

$$U_i(f) = \pi_i(f) + \rho_i \sum_{\substack{n \rightarrow f \\ n \in N_i}} \varphi(n)\sigma(n, f) \tag{7}$$

According to Definition 3 player i's utility in the reciprocity game is the sum of the following two terms: The first summand is simply player i's **material payoff** $\pi_i(f)$. The second summand - which we call **reciprocity utility** - is composed of:

- The positive constant ρ_i, the **reciprocity parameter**. This constant is an individual parameter which captures the strength of player i's reciprocal preferences. It takes account of the fact that individuals differ in their reciprocal inclination. The higher ρ_i, the more important is the reciprocity utility as compared to the utility arising from the material payoff. Note that if ρ_i equals zero, player i's utility is equal to his material payoff, i.e., identical to the standard game. If, in addition, ρ_j also equals zero, the reciprocity game collapses into the standard game.

- The **kindness term** $\varphi(n)$ which measures the kindness player i experiences from j's (expected) actions. As noted above it is positive if player j is considered as kind and negative if player j is considered as unkind. The kindness term comprises outcomes and intentions.

- The **reciprocation term** $\sigma(n, f)$. It measures the effect of the reciprocal action.

- The product of the *kindness* and the *reciprocation term* measures the reciprocity utility in a particular node. If the kindness term in a particular node n is greater than zero, player i can *ceteris paribus* increase his utility if he chooses an action in that node which increases player j's payoff. The opposite holds if the kindness term is negative. In this case, player i has an incentive to reduce player j's payoff.

- Since kindness is measured in each node where player i has the move, the overall **reciprocity utility** is the sum of the reciprocity utility in all nodes (before the considered end node), weighted with the reciprocity parameter.

3.4 The Reciprocity Equilibrium

The introduced preferences form a psychological game (Geanakoplos, Pearce and Stacchetti (1989)). In psychological games, the utility of a player i does not only depend on the selected strategies of the players but also on the beliefs player i has (compare Definition 3). Note, however, that beliefs are not part of the action space. Put differently, beliefs cannot be formed strategically, i.e., they are taken as given. Given the beliefs, player i chooses his optimal strategy. The additional requirement in a psychological Nash equilibrium as compared to a Nash equilibrium is that all

beliefs match actual behavior. This means, that an optimal strategy is only part of an equilibrium if the beliefs are also consistent with the actual behavior.[11]

Geanakoplos, Pearce and Stacchetti (1989) show that the refinement concept of subgame perfectness can also be applied to psychological Nash equilibria. In our reciprocity game we will call a subgame perfect psychological Nash equilibrium a **reciprocity equilibrium**. If $\rho_i = \rho_j = 0$, the definition of a reciprocity equilibrium is equivalent to the definition of a subgame perfect Nash equilibrium.[12]

The concept of psychological game theory allows us to formalize the concept of intentions in a straightforward way. As we have seen in Section 2 and as we will show in the next section, the formalization of intentions is important since purely consequentialistic models fail to explain relevant aspects of the data.

4 Applications

Many scholars have presented compelling evidence on the relevance of reciprocity in a variety of economically important situations. The most clear-cut evidence on positive and negative reciprocity stems from controlled laboratory experiments. In this section we, therefore, discuss the predictions of our theory in different experimental games. The games we study are the ultimatum game, the gift-exchange game, a reduced best-shot game, market games with proposer or responder competition, the dictator game, the prisoner's dilemma, public goods games, and the investment game. Notice that in all figures which illustrate our propositions we use the *same* set of parameters ($0 \leq \rho_i \leq 4$ or $\rho_i = 2$, $\varepsilon_i = 0.2$).

4.1 Negative reciprocity: The ultimatum game

Let us start our applications section with an example of negative reciprocity. The most known game in which negative reciprocity applies is the ultimatum game. In this two person sequential move game, the first mover ("proposer") is allocated an

[11] Assuming rationality of players guarantees that the optimal strategy is chosen. The coincidence of actual behavior and beliefs, on the other hand, cannot simply be justified on the basis of rationality. It is very plausible, however, to assume that possible inconsistencies between actual behavior and first order beliefs will die out with some experience. Let us therefore assume that the first order belief of player j (eventually) matches actual behavior of player i. Now, if player i forms his second order belief in the same way as player j forms his first order belief (e.g., by observing actual behavior or introspection) then player i will be able to match his second order beliefs with the first order beliefs of player j.

[12] A remark on the existence of reciprocity equilibria: In the presented form, a reciprocity equilibrium does not always exist because the function Ω is discontinuous at $\pi_i = \pi_i^0$. A minor technical modification of Ω, however, guarantees the existence of a reciprocity equilibrium. For the ease of exposition we delegate the existence proof for the modified Ω to Appendix 1.

amount of money (which we normalize to 1). The proposer has to divide this amount between himself and a second mover ("responder"). He may offer any feasible amount c to the responder, i.e., $0 \leq c \leq 1$. After the offer is revealed to the responder, the latter can either accept or reject it. If she accepts, the resulting payoffs are $1-c$ for the proposer and c for the responder. If the responder rejects the offer, payoffs are zero for both parties. Given the standard assumptions, the outcome according to the subgame perfect Nash equilibrium is ($c = 0$; accept).

The ultimatum game has intensively been studied. Overviews of experimental results are presented, e.g., in Güth, Schmittberger and Schwarze (1982), Thaler (1988), Güth (1995), Camerer and Thaler (1995) and Roth (1995).[13] The reported behavioral regularities are quite robust and can be summarized as follows: There are (i) practically no offers that exceed 0.5, (ii) the modal offers lie in a range between 0.4 and 0.5, (iii) offers below 0.2 are extremely rare, and (iv) whereas offers close to 0.5 are practically never rejected, the rejection rate for offers below 0.2 is rather high. Thus, the standard subgame perfect prediction is strongly refuted by the stylized facts.

We now state our predictions. Upon acceptance, material payoffs are $1-c$ for the proposer and c for the responder, respectively. Let p denote the probability that the responder accepts the offer.

Notation: For the following applications it is useful to introduce the notations $[c]_{[a}^{b]} := \min(b, \max(a, c))$, $[c]_{[a} := \max(a, c)$, and $[c]^{b]} := \min(b, c)$. The expression $[c]_{[a}^{b]}$ is simply c if c lies within the interval $[a, b]$ and is the nearest end point of $[a, b]$ otherwise.

Proposition 1 *If ρ_1 and ρ_2 are positive there is a unique reciprocity equilibrium (c^*, p^*) in the ultimatum game as follows:*

$$p^* = \begin{cases} \left[\frac{c}{\rho_2 \cdot (1-2c)(1-c)}\right]^{1]} & if\ c < \frac{1}{2} \\ 1 & if\ c \geq \frac{1}{2} \end{cases} \tag{8}$$

$$c^* = \max\left[\frac{1 + 3\rho_2 - \sqrt{1 + 6\rho_2 + \rho_2^2}}{4\rho_2}, \frac{1}{2} \cdot (1 - \frac{1}{\rho_1})\right] \tag{9}$$

If either ρ_1 or ρ_2 is zero p^* and c^* are the limits of the above formulas where ρ_1 and ρ_2 approach zero from above.

If ρ_1 and ρ_2 are both zero, $p^* = 1$ and $c^* = 0$.

[13] An interesting extension of the ultimatum game to more than two players is presented in Güth, Huck and Ockenfels (1996).

The proof of Proposition 1 is given in Appendix 3.

Discussion: Let us first look at the responder's behavior. Equation (8) reveals the conditions that determine the acceptance probability p^* of an offer c in the reciprocity equilibrium: First, if the proposer's offer is equal or higher than half of the pie, i.e., $c \geq \frac{1}{2}$, the responder will *always* accept the offer. This holds independent on the responder's concern for reciprocity (compare the second row of Equation (8)). Second, for offers smaller than half of the pie the willingness to accept an offer is increasing in the level of the offer and decreasing in the responder's concern for reciprocity, ρ_2 (see the first row of Equation (8)). Figure 2 depicts the acceptance probability as a function of the level of the offer for a given reciprocal inclination of the responder. The figure neatly captures the essence of negative reciprocity: The lower the offer, the higher is the willingness of a reciprocally motivated responder to punish the proposer by rejecting the offer. Note that as ρ_2 gets higher (lower) the acceptance-curve shifts to the right (left).

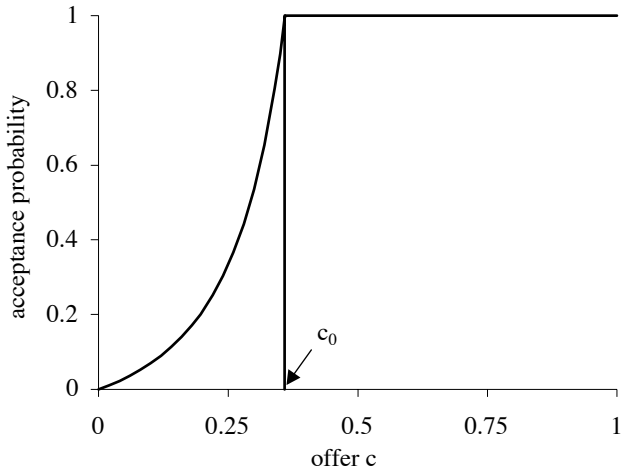

Figure 2: Acceptance probability in the ultimatum game dependent on the offer of the proposer for a given reciprocity parameter of the responder of $\rho_2 = 2$.

We now turn to the proposer. Equation (9) shows that his equilibrium choice of c depends on two expressions. Whereas the first expression depends on the *responder's* reciprocal inclination, i.e., on ρ_2, the second expression depends on the *proposer's* concern for reciprocity, i.e., on ρ_1. The second expression the proposer's *intrinsic* concern for an equitable outcome. If he is an egoistic player the expression is zero. If ρ_1 is large, however, he will offer a positive c. The first expression can be interpreted as an *extrinsic* constraint to offer a positive c: This expression exactly defines the smallest offer that just guarantees an acceptance probability of 1. (We call this

offer c_0, i.e., $c_0 := \frac{1+3\rho_2-\sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}$, compare Figure 2). Obviously, the responder's concern for reciprocity is crucial for the value of c_0. For example, if the responder is a selfish player (with $\rho_2 = 0$) the value of c_0 is equal to zero. The higher player 2's concern for reciprocity, the higher the value of c_0. As ρ_2 gets very large, c_0 approaches $\frac{1}{2}$. The dependence of c_0 on ρ_2 is shown in Figure 3. It is remarkable that even for small values of ρ_2 the proposer is 'forced' to make a rather generous offer in order to avoid a rejection.

The equilibrium offer c^* is the maximum of the first and the second expression of Equation (9): This means, e.g., that in case a selfish proposer plays against a reciprocal responder he will offer the more, the higher ρ_2. In this case the extrinsic constraint is binding. If, however, the responder has a very low ρ_2, i.e., he accepts practically any offer, the equilibrium offer is determined by the proposer's concern for an equitable outcome: If he is rather selfish too, his offer will be (close to) zero. If, on the other hand, ρ_1 is rather high, the offer will be high as well.[14]

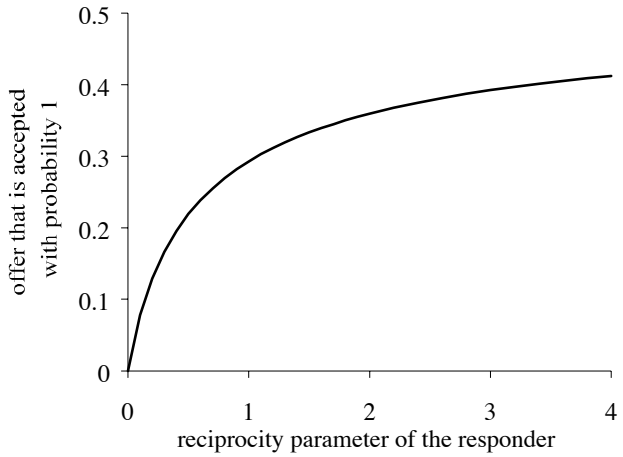

Figure 3: Offers in the ultimatum game that guarantee an acceptance with probability 1, dependent on the the reciprocity parameter of the responder (ρ_2).

4.1.1 The role of intentions

Before we turn to the next game we will briefly discuss the predictions of our theory for the non-intentional treatment reported by Blount (1995). In her treatment the proposers' offers were not chosen by human subjects but instead randomly selected. Consequently, a low offer did not signal any (bad) intentions. As the data of Blount's

[14] The range of ρ_1- and ρ_2- combinations where the equilibrium offer c^* equals c_0 is given by $\rho_2 \geq \frac{\rho_1(\rho_1-1)}{\rho_1+1}$. This holds in particular if $\rho_2 \geq \rho_1$.

experiment reveals, the acceptance rate for a given offer is *much higher* than in the 'regular' treatment. However, even in the absence of intentions some subjects reject extremely disadvantageous offers. Our theory predicts exactly these two stylized facts. In the non-intentional treatment the equilibrium acceptance rate for p^* is given by:

$$p^* = \begin{cases} \left[\frac{c}{\varepsilon_2\rho_2\cdot(1-2c)(1-c)}\right]^{1]} & \text{if } c < \frac{1}{2} \\ 1 & \text{if } c \geq \frac{1}{2} \end{cases}$$

Figure 4 depicts the predicted acceptance probabilities in the 'regular' ultimatum game and in Blount's treatment for a given ρ_2. The upper graph corresponds to the Blount-experiment whereas the lower graph shows the acceptance behavior in the 'regular' treatment. As can be seen in the figure, a responder's acceptance probability for low offers is higher if intentions are absent.

The lower the outcome concern parameter ε_2, the more the upper graph shifts to the left. Put differently, the more a responder is concerned with intentions relative to outcomes, the more likely she will accept very low offers in Blount's non-intentional treatment. On the other hand, if a responder is purely outcome oriented, i.e., if $\varepsilon_2 = 1$ holds, she exhibits the *same* behavioral pattern as in the 'regular' treatment. As the data reveals, however, people care about intentions (i.e., $\varepsilon_i < 1$).

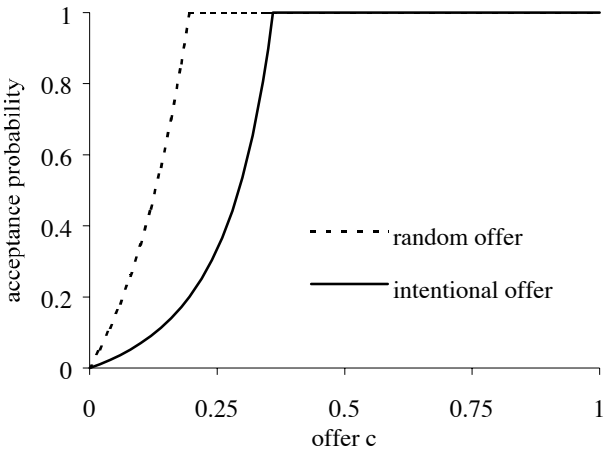

Figure 4: Acceptance probabilities in the ultimatum game with intentions (lower graph) and without intentions (upper graph) dependent on the offer. The parameters $\rho_2 = 2$ and $\varepsilon_2 = 0.2$ are given.

4.2 Positive reciprocity: The gift-exchange game

Many experiments have demonstrated the importance of positive reciprocity. One well known experiment is the gift-exchange game. In this two-person sequential game, the first mover (called an employer) offers a wage w to a second mover (called a worker). The worker can either take it or leave it (see e.g., FEHR, GÄCHTER AND KIRCHSTEIGER (1997)). If the worker rejects, both earn nothing. If the worker takes the wage offer, she has to make an effort decision e. Providing effort above the minimum effort level is costly with $c(e)$ being a convex effort cost function. Payoff functions are given by $\pi_1 = ve - w$ for employers and $\pi_2 = w - c(e)$ for workers, respectively. A rational and selfish worker will - irrespective of the wage paid - choose the minimum effort level. With backward induction, the employer will offer only the lowest possible wage. Contrary to this prediction the main experimental findings are (i) that wages clearly exceed the lowest possible wage and (ii) that there is a positive wage-effort relation. Both findings are remarkably robust. They hold in bilateral institutions as well as in competitive market institutions (see, e.g., FEHR, KIRCHSTEIGER AND RIEDL (1993), FEHR AND FALK (1999) and GÄCHTER AND FALK (1997)).

In our analysis we restrict $w \in [0,1]$, $e \in [0,1]$, we normalize v to 1 and write the convex effort cost function as $c(e) = \alpha e^2$ with $\alpha \leq \frac{1}{4}$. To ease the proposition of the equilibria in the gift-exchange game, we use the following definitions. First, let $\widetilde{w}(\alpha, \rho_1, \rho_2)$ be the wage that would be chosen by the employer if w and e would not be restricted to be smaller than 1. The exact formula for $\widetilde{w}(\alpha, \rho_1, \rho_2)$ is given in the proof of the proposition. We further define $\overline{w}(\alpha, \rho_2) = \frac{1+a}{2} + \frac{a}{\rho_2}$ which is the minimal wage that guarantees an effort choice of one. The following proposition now shows that the stylized facts of the gift-exchange game are replicated by our theory:

Proposition 2 *In the gift-exchange game there are the following reciprocity equilibria* (w^*, e^*)*:*

(i) If $\rho_2 = 0$*, then there exists a unique equilibrium:*

$$w^* = e^* = 0 \tag{10}$$

(ii) If $\rho_2 > 0$*, then*

$$e^* = \left[\frac{-2\alpha - \rho_2 + \sqrt{(2\alpha + \rho_2)^2 + 8\alpha\rho_2^2 w}}{2\alpha\rho_2} \right]_{]0}^{1]} \tag{11}$$

There is always an equilibrium given by:

$$w^* = \left[\min\left(\overline{w}(\alpha, \rho_2), \widetilde{w}(\alpha, \rho_1, \rho_2)\right) \right]_{]0}^{1]} \tag{12}$$

If $\varepsilon_1\rho_1 \leq \frac{\rho_2(-\rho_2+2\alpha+2\alpha\rho_2)}{2\alpha(-2\alpha-\rho_2+2\alpha\rho_2)}$ *and* $\overline{w}(\alpha,\rho_2) \leq 1$, *then there is a second equilibrium which is given by:*

$$w^* = \overline{w}(\alpha,\rho_2)$$

The proof of Proposition 2 is given in Appendix 3.

Discussion: We illustrate Proposition 2 with the help of Figures 5 to 7. In Figure 5 we depict how - in equilibrium - a worker's effort choice depends on the wage paid by the employer given her reciprocal motivation, ρ_2 (compare equation (11) of Proposition 2). This figure neatly captures the essence of *positive reciprocity* as reported in the gift-exchange experiments. Reciprocal workers provide higher effort levels the higher the wage paid by 'their' employers. Moreover, workers will - for a given wage - choose higher effort levels, the stronger their reciprocal inclination (compare equation 11). This relationship is depicted in Figure 6. The upper graph shows the equilibrium effort choice for a wage $w = 0.5$, whereas the lower depicts the optimal effort for a lower wage ($w = 0.3$). Note that the effort level is zero if and only if $\rho_2 = 0$ or $w = 0$.

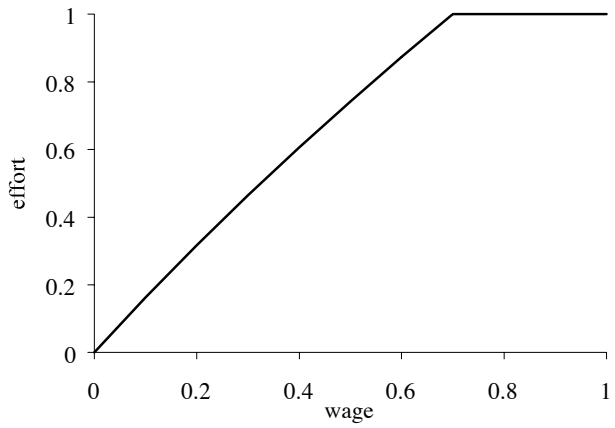

Figure 5: Effort choice dependent on the wage paid for a given ρ_2 and a given α ($\rho_2 = 2$ and $\alpha = 0.2$).

How can we characterize an employer's optimal choice, given a worker's effort behavior? In Figure 7 we show how the equilibrium wage depends on ρ_2 (for a given ρ_1). In equilibrium, an employer will - for a small ρ_2 - offer a wage equal to zero which also implies an effort choice of zero. If, however, the worker is sufficiently reciprocal the wage offer increases in ρ_2.[15] Notice in Figure 7 that this increase stops as ρ_2

[15] As can be seen in Figure 7 there is an intermediate range of ρ_2 where two equilibria exist. The equilibrium corresponding to the upper branch is described below.

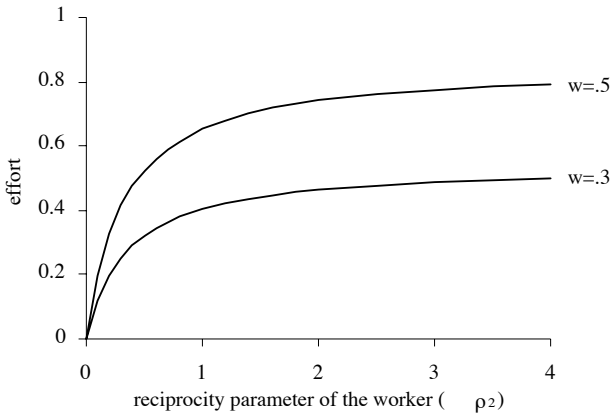

Figure 6: Equilibrium effort choice dependent on the worker's reciprocity parameter ρ_2 a given α and two wage levels ($\alpha = .2$, $w = .3$ and $w = .5$).

reaches a particular level. At this level the employer gets the *maximal* effort level. Thus, it makes no sense to pay a higher wage, neither for a reciprocal nor for a selfish employer. To the contrary, as ρ_2 gets higher the lowest wage that guarantees the maximal effort level actually decreases.

How does the employer's concern for reciprocity ρ_1 influence its decision? To answer this question notice first that the worker can always ensure at least an equal share for herself, i.e., the employer is always in a disadvantageous situation. Whereas a sufficiently selfish employer does not bother about this, a reciprocal employer does. Consequently, for small ρ_1 an employer pays the profit maximizing wage. As ρ_1 gets higher, however, the employer actually pays a *lower* wage - due to the impact of negative reciprocity.

Figure 7 reveals another interesting feature of Proposition 2: For sufficiently high ρ_1 there is an intermediate range of ρ_2 where *two* equilibrium wages exist. In the equilibrium corresponding to the upper branch the employer pays a wage that guarantees an effort of 1. This equilibrium can be described as follows: The employer expects to get an effort of 1, i.e., the *maximum* effort level. This means that the worker could *not* have chosen a higher effort level. Thus, even though the employer is in a disadvantageous situation he cannot infer that the worker *wanted* to treat the employer unkindly. As a consequence the employer does not have the desire to punish the worker and therefore chooses the payoff maximizing wage, i.e., a rather high wage that guarantees an effort level of 1. In the equilibrium corresponding to the lower branch in Figure 7, the employer expects that the worker provides an effort strictly *below* 1. This equilibrium may coexist with the former for the following reason: The

fact that the worker chooses an effort smaller than 1 is considered as very unkind since she *could have chosen* a (slightly) higher effort level. Therefore, the (reciprocal) employer has an incentive to punish the worker, i.e., to pay a low wage. This low wage now induces the worker to provide the low effort level the employer expected. In a certain sense the employer's equilibrium behavior can be understood as a kind of "self fulfilling prophecy": If the employer believes that the worker is nice, the worker will indeed *be* nice. If, on the other hand, the employer believes that the worker is mean, the worker will in fact behave accordingly.

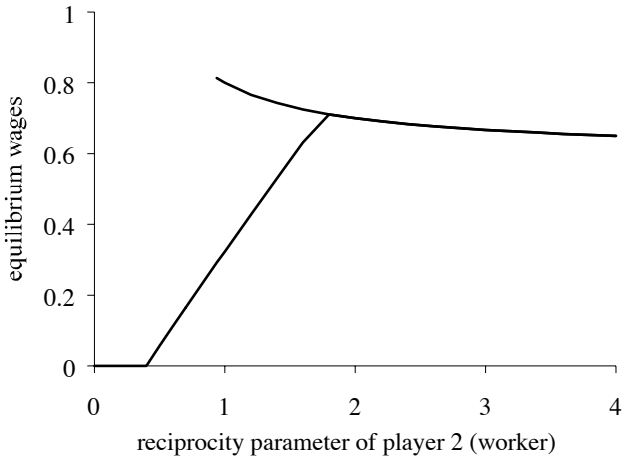

Figure 7: Equilibrium wage offer dependent on ρ_2 for a given ρ_1 and α ($\rho_1 = 2$ and $\alpha = 0.2$). The lower branch corresponds to equilibria with $e^* < 1$ whereas in the upper branch $e^* = 1$ holds.

4.2.1 The role of intentions

Concluding our discussion of the gift-exchange game we present our predictions for the treatment reported by Charness (1996). In his experiment wages are not determined by the person the worker is interacting with. Rather, a third party or a random mechanism determines the wage. Compared to the 'regular' treatment, this leads to a *weaker* correlation between wages and effort levels. The reason is that in Charness' treatment a high wage does not signal any (kind) intentions. Whereas purely outcome oriented models cannot distinguish between these two treatments, our theory does. For a given reciprocal motivation ρ_2 we predict a weaker correlation between wages and effort levels. This means that the *same* worker will supply a *lower* effort for a particular received wage in the random-treatment compared to the 'regular' treatment. Figure 8 shows that result.

Figure 8: Equilibrium effort dependent on the wage. In the upper graph, the wage is chosen by a human being, in the lower graph a random wage is given. ($\varepsilon_2 = 0.2$, $\alpha = 0.2$)

4.3 Identical outcomes yield different responses: A comparison between best-shot and ultimatum games

The best-shot game was introduced by Harrison and Hirshleifer (1989) and Prasnikar and Roth (1992). The interesting feature of this experiment is that second movers are willing to accept a higher degree of inequity than in the ultimatum game results. This difference cannot be explained by the inequity aversion models by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). Our explanation for the difference between behavior in the best-shot game and the ultimatum game rests on the importance of intention. This will become clear with the help of the following two games.

In Figure 9 we present a reduced best-shot game which captures the key aspects of the original game. The crucial feature of this game (and the richer original game) is that the first mover can only offer a payoff share that is either very advantageous or very disadvantageous to himself (8/2 or 2/8). This is different in the reduced ultimatum game where - alternative to the very advantageous offer (8/2) - the first mover can also choose the offer (5/5).

As the following proposition reveals, our theory predicts a higher acceptance probability for the unkind offer (8/2) in the reduced best-shot game, compared to the reduced ultimatum game.

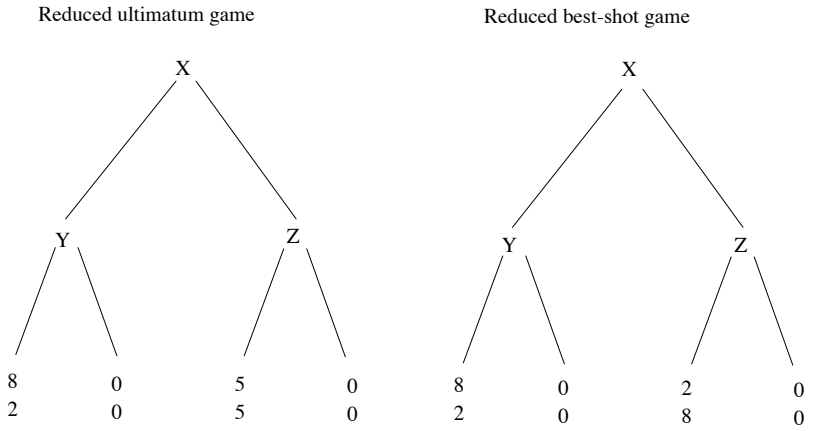

Figure 9: The game trees of a reduced ultimatum game and a reduced best-shot game.

Proposition 3 *In the normalized*[16] *reduced ultimatum game there is a unique reciprocity equilibrium. It is given by:*

$$r^* = 1 \tag{13}$$

$$q^* = \left[\frac{5}{12}\frac{1}{\rho_2}\right]^{1]} \tag{14}$$

$$p^* = \left[\frac{0.8q^* - 0.5}{0.6(0.5 - 0.2q^*)q^*\rho_1}\right]^{1]}_{[0} \tag{15}$$

If $\rho_2\varepsilon_2 < 1\frac{2}{3}$ *there is a unique reciprocity equilibrium in the reduced best-shot game which is given by:*[17]

$$r^* = 1 \tag{16}$$

$$q^* = \left[\frac{5}{12}\frac{1}{\varepsilon_2\rho_2}\right]^{1]} \tag{17}$$

$$p^* = \left[\frac{1 + \frac{0.8q^* - 0.2}{0.6(0.8 - 0.2q^*)\rho_1}}{1 + q^*}\right]^{1]}_{[0} \tag{18}$$

If ρ_1 *or* ρ_2 *are zero,* p^* *and* q^* *are the limits of the above formulas where* ρ_1 *and* ρ_2 *approach zero from above.*

[16] The payoffs shown in the game tree of Figure 1 are normalized with a factor of $\frac{1}{10}$.

[17] In Appendix 2 we describe also the equilibria for $\rho_2\varepsilon_2 \geq 1\frac{2}{3}$. This is omitted here for three reasons. First, because this case is rather complicated. Second, in all figures where we show the predictions of our model, we use the *same* parameter set ($\varepsilon_i = 0.2$ and $\rho_i \in [0, 4]$) which we consider as the relevant range for ε_i and ρ_i. For $\rho_2\varepsilon_2 \geq 1\frac{2}{3}$ we are far outside of this range. Third, we are mostly interested in the comparison of second movers' behavior in the best-shot and the ultimatum game and Equation 17 also holds for $\rho_2\varepsilon_2 \geq 1\frac{2}{3}$.

The proof of Proposition 3 is given in Appendix 3.

Discussion: In our discussion we focus on the second mover behavior. First, note that the kind offer (2/8) is always accepted in the best-shot as well as in the ultimatum game ($r^* = 1$). The acceptance probability q^* of the unkind offer (8/2) differs, though. This difference is shown in Figure 10.

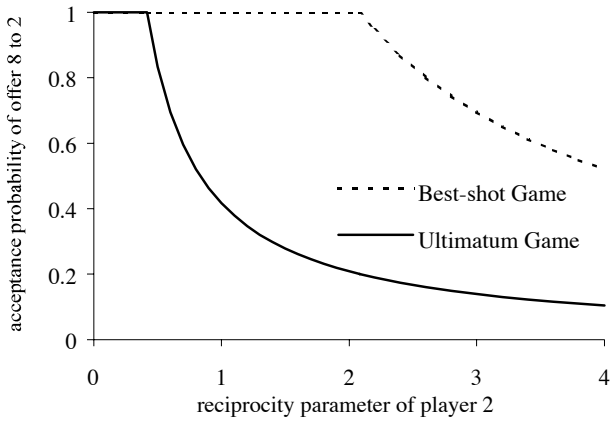

Figure 10: The acceptance probability if the unkind offer (8,2) is higher in the reduced best-shot game (upper graph) than in the reduced ultimatum game (lower graph).

The upper graph shows the acceptance probability of the unkind offer in the reduced best-shot game, the lower shows that of the reduced ultimatum game. In both games, we see that the acceptance probability of the unkind offer decreases in player 2's concern for reciprocity, i.e., in ρ_2. We also see, however, that for a given ρ_2 the acceptance probability is lower in the ultimatum game compared to the best-shot game. Thus, in the *best-shot game a reciprocal second mover is willing to accept a higher degree of inequity.*

The reason is that in the best-shot game the only alternative to offering (8/2) is to offer a very disadvantageous offer, namely (2/8). Since such an alternative implies that the first mover switches from an advantageous to a disadvantageous situation, player 2 will not find it very unkind if he gets the 'unkind' offer, because she cannot infer very bad intentions from this offer. Things are quite different in the reduced ultimatum game. Here, the first mover has the alternative to choose an equitable outcome. Consequently, it *does* signal bad intentions and is considered as quite unkind if this opportunity is not chosen. Thus, dependent on the first mover's alternatives the *same* offer will be accepted with a different probability. This difference is confirmed in the experimental study by Falk, Fehr, and Fischbacher (1999): The rejection

rate of the (8/2)-offer is 27 percent in the reduced best-shot game and 44 percent in the reduced ultimatum game.

4.4 Competition

In the preceding games we have analyzed only bilateral interactions. In particular, we restricted our analysis to games without any competition at all. In this section we, therefore, apply our theory to n-person games and show how competitive pressure interacts with reciprocal preferences.

It is a well established fact in experimental economics that in competitive institutions market outcomes converge very well towards the outcome predicted by standard economic theory (Smith (1982), Davis and Holt (1993)). This holds even in markets where the equilibrium outcome is very "unfair" in the sense that almost the whole surplus is reaped by one side of the market. Put differently, it seems that in a competitive environment subjects behave as if they were completely selfish. How can a theory of reciprocity possibly account for this fact? In this section we show that our theory can explain not only why in bilateral institutions outcomes tend to be "fair" but also why in markets reciprocal subjects' behavior gives rise to equilibria that imply an extremely uneven distribution of the gains from trade.

As an illustration we analyze a market game with proposer competition which has been studied by Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991). In this game there are $n-1$ proposers who simultaneously propose an offer $c_i \in [0,1]$ to the responder with $i \in [1, n-1]$. These offers are revealed to the responder who has to decide whether to accept or reject the highest offer c_{max}. If more than one proposer offers c_{max} a random mechanism determines whose offer will be selected. Payoffs are exactly as in the ultimatum game, i.e., the proposer whose offer is accepted receives $1-c_{max}$ and the responder gets c_{max}. A proposer whose offer is not accepted receives a payoff of zero. If the responder rejects c_{max}, all receive nothing. Assuming rational and purely selfish players the subgame perfect outcome is straightforward: At least two proposers offer $c_{max} = 1$ which is accepted by the responder. In fact, this prediction is supported by the experimental data. Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991) report that in their experimental sessions (with 9 proposers) this equilibrium outcome was reached after a few periods. Moreover, these results were very robust and showed up in four different countries. Given the experimental results of the preceding bilateral games this is remarkable since in this equilibrium the proposers earn practically nothing while the responder gets the whole surplus. As it turns out, the prediction of our theory coincides with the standard prediction.

Proposition 4 *In a reciprocity equilibrium in the market game with proposer competition at least 2 proposers offer* $c_{max} = 1$ *which is accepted by the responder.*

The proof of Proposition 4 is given in Appendix 3. It requires an extension of our theory to n-person games. This extension is simply notational and will be discussed in Appendix 2.

Discussion: The striking feature of Proposition 4 is that - independent of their reciprocal inclination - proposers will accept a very uneven distribution of the pie. The intuition of that result is that in a competitive market a proposer has *no chance* to achieve a "fair" outcome: From the *ultimatum game* we know that a responder will accept any offer larger or equal to 0.5 with certainty. Since offering more decreases a proposer's material and reciprocity utility a proposer will - in the ultimatum game - never offer more than that. Assume that in a market game with two proposers a reciprocal proposer i refuses to offer more than 0.5. What will the other proposer do? By infinitesimally overbidding player i's offer he can on the one hand increase his material payoff by a positive amount (because he can increase the winning probability from $\frac{1}{n-1}$ to 1). On the other hand, the reciprocity disutility resulting from the unfair relation to the responder does only change infinitesimally. This means that player i's refusal to propose more than 0.5 is not an effective tool to achieve a "fair" outcome. As a consequence, he tries to overbid the other proposer to get at least a minimal share of the pie. This "overbidding", however, leads inevitably to the equilibrium described in Proposition 4.

Another experiment which yields an "unfair" outcome is the market game with **responder competition** (Güth, Marchand, and Rullière (1997)). In this game one proposer faces n responders, who may "underbid" each other in accepting low offers. As in the market game with proposer competition, our theory is compatible with the very "unfair" outcome predicted by standard economic theory. According to the latter prediction and the stylized experimental facts responders accept much lower offers than in the ultimatum game. This holds even if all responders are reciprocally motivated. The intuition is similar to the market game with proposer competition. We saw that in the *ultimatum game* a responder can "force" the proposer to make a "fair" offer by credibly threatening to reject very low offers. This punishment device is ineffective in the market game with responder competition if there are other responders who accept low offers.

Thus our theory does not only predict "fair" outcomes. Rather, it correctly explains "out-of-sample-facts", i.e., very "unfair" equilibria that occur in competitive institutions.

4.5 Further games

In this section we briefly discuss the intuition of our theory's predictions in some further games. The propositions and proofs have rigorously been developed in a previous version.[18] In this section we report only the main results.

4.5.1 The dictator game

The dictator game is a very simple two person game. The task of the first mover (the so-called "dictator") is to divide an amount of money between himself and a counterpart (the "receiver"). Let 1 be the amount of money and c the share for the receiver. The dictator is free to choose any feasible division he wants to ($0 \leq c \leq 1$). The receiver has no choice to make, i.e., she has to accept any amount sent to her. The payoff for the receiver is simply the amount c she has been sent by the dictator. The dictator's payoff is given by the residual amount $1 - c$.

The dictator game has been studied, e.g., by Forsythe, Horowitz, Savin and Sefton (1994) and by Hoffman, McCabe, Shachat and Smith (1994) and Eckel and Grossman (1998). The stylized facts can be summarized as follows. (i) Dictator offers larger than half of the pie, i.e., $c > \frac{1}{2}$ are practically never observed. (ii) Roughly 80 percent of the offers are between zero and half of the pie, i.e., $0 < c \leq \frac{1}{2}$. However, *compared to the ultimatum game, the distribution of offers is shifted towards zero.* (iii) About 20 percent of the dictators offer the amount predicted by standard game theory, i.e., they offer exactly zero.[19]

Compatible with these stylized facts our theory predicts a unique reciprocity equilibrium. In this equilibrium the dictator offers $c^* = \left[\frac{1}{2} \cdot \left(1 - \frac{1}{\varepsilon_1 \rho_1}\right)\right]_{[0}$.[20] Thus, a dictator's proposal depends (i) on ρ_1, i.e., on how strong his other-regarding preferences enter his utility and (ii) on his pure outcome concern parameter ε_1. If $\varepsilon_1 \rho_1 > 1$, the dictator offers a positive amount of money. Even for a very high values of $\varepsilon_1 \rho_1$, however, the dictator's offer will never exceed $\frac{1}{2}$, i.e., he will never offer more than he keeps for himself. If $\varepsilon_1 \rho_1 \leq 1$, the dictator chooses an offer equal to zero. Comparing the equilibrium offers in the dictator game with those of the ultimatum game, we see that the equation that determines the equilibrium offer in the dictator game equals the second expression in equation (9) of Proposition 1 - if we replace ρ_1 by $\varepsilon_1 \rho_1$. Since $\varepsilon_1 \leq 1$, the same person will always offer at least as much in the ultimatum game as

[18] This version, including propositions and proofs is available on request.

[19] It should be noted that especially the results of the dictator game are not very robust with respect to treatment variations. Increasing, e.g., the social distance among participants of an experiment and the experimenter (double blind treatment) increases the percentage of zero proposals (compare Hoffman, McCabe and Smith (1996)).

[20] This holds, if $\varepsilon_1 \rho_1 > 0$. If $\varepsilon_1 \rho_1$ equals to zero, c^* is chosen equal to zero.

in the dictator game.

Thus, consistent with stylized facts (i) to (iii), our theory predicts that dictators offer between zero and half of the pie and that the distribution of offers in the dictator game is shifted downwards compared to the corresponding distribution in the ultimatum game.

4.5.2 The Prisoner's Dilemma and Public Goods Games

The sequential prisoner's dilemma consists of two stages. At the first stage, player 1 can either cooperate or defect. After observing player 1's choice player 2 also chooses either to cooperate or to defect. Assuming rational and selfish actors, both players have a dominant strategy not to cooperate.

Contrary to this prediction, our theory predicts the following: First, player 2 always defects if player 1 has defected beforehand. This means that according to our theory we will find no altruistic, i.e., non-conditional cooperation. Second, if player 2 is sufficiently reciprocally motivated, there is a positive probability that player 2 reacts to cooperation with cooperation. Thus, our theory predicts *conditional cooperation* if player 2's concern for reciprocity is strong enough.

Experimental studies of sequential versions of the prisoner' s dilemma are reported in Bolle and Ockenfels (1990) and Clark and Sefton (1998). The results of their studies are in line with our predictions. In particular, unconditional cooperation is practically inexistent. In their conclusion Clark and Sefton, e.g., note that "cooperation is better regarded as reciprocation rather than unconditional altruism: second-movers cooperate quite frequently in response to cooperation by first-movers, but rarely cooperate after the first-mover defects" (p. 17).

We also analyze the simultaneous prisoner's dilemma. According to our theory, we get less cooperation if players choose simultaneously compared to the sequential move structure. There is strong evidence in favor of this prediction. Experimental studies by Watabe, Terai, Hayashi, and Yamagishi (1996) and Hayashi, Ostrom, Walker, and Yamagishi (1998) show that cooperation rates among first movers in the sequential prisoner's dilemma are much higher, compared to simultaneous move games.

The strategic structure of the prisoner's dilemma is very similar to that of *public goods games*. The major difference is that in a public goods game players have more strategies than just to cooperate or to defect. In a typical public goods game each player is provided with an endowment of 20 tokens, say, that can be allocated to a public good or that can be kept for private consumption. The marginal per capita return of an investment into the public good is usually smaller than one, i.e., providing zero tokens to the public good is a dominant strategy (whereas it is in the common

interest to fully cooperate). Most public goods experiments have been conducted as simultaneous move games. However, there is a recent experiment by FISCHBACHER, GÄCHTER AND FEHR (1999) where subjects could *conditionally* indicate how many tokens they wanted to invest into the group account. Despite the fact that the best reply is to provide zero tokens independent of the other group members' contributions, subjects on average contributed more the higher the contributions of the other group members. This 'conditional cooperation strategy' was in most cases specified such that subjects provided less than the group average. This is exactly what our theory predicts: In a public goods game subjects with a sufficiently high reciprocal inclination will conditionally cooperate but always cooperate slightly less than the other player(s). Moreover, our theory replicates another stylized fact of public goods games, namely that the propensity to cooperate increases in the marginal per capita return of an investment into the public good (see LEDYARD (1995)).

4.5.3 The Investment Game

In this game, first movers are endowed with an amount of money which they can keep or transfer to the second mover. Any amount y transferred is tripled. The second mover then decides on a countertransfer z, with $0 \leq z \leq 3y$. If both players are rational and selfish money maximizers the predicted outcome is $y = z = 0$.

The investment game was studied by BERG, DICKHAUT AND MCCABE (1995) and MILLER (1997). The authors report that contrary to the standard prediction, first movers on average transfer positive amounts of money and do receive positive countertransfers. Our theory predicts exactly this: If second movers are sufficiently reciprocal, they make conditional countertransfers, i.e., z increases in y. This holds the more, the greater the second mover's reciprocal inclination. First movers choose positive transfers if the expected countertransfer is sufficiently high, i.e., if second movers are sufficiently reciprocal.

5 Summary

In this paper, we present a formal theory of reciprocity where *reciprocity* means the behavioral response to perceived kindness or unkindness. We argue that the perceived kindness of an action is composed of two elements, namely the *outcome* that results from the action and the action's underlying *intentions*. In our theory we implement an *equity* based reference standard for the evaluation of the kindness or unkindness of the chosen outcome. This outcome, however, will be perceived differently depending on the intentions involved. If intentions are absent, the behavioral response will be less intense. However, even in this case, subjects still experience the *outcome per se*

as either advantageous or disadvantageous and shows some reciprocal response.

Our theory predicts the stylized facts of a wide variety of experimental games: In the *ultimatum game*, proposers offer between zero and half of the total pie. Rejections are decreasing in the level of the offer and increasing in the strength of the responder's concern for reciprocity. In the *dictator game*, the theory predicts offers which are lower than in the ultimatum game. We analyze a *reduced best-shot game* and show that - as in the experimental data - subjects are willing to bear a higher degree of inequity than in the ultimatum game. In the *gift-exchange game* our theory predicts a positive relationship between wages and effort levels. Moreover, firms offer above-minimum wages - as reported in the experimental literature. In the *sequential prisoner's dilemma* we predict conditional cooperation. Similarly, in *public goods games* subjects contribute more, the more they expect others to contribute. Moreover, contributions increase in the marginal per capita return of an investment to the public good. In the *investment game* we find a positive relation between the first mover's transfer and the second mover's countertransfer. In an n-person *market game with proposer competition* the theory predicts that at least two proposers offer the total pie to the responder who accepts this offer. Thus, the whole surplus is reaped by the responder. This extremely "unfair" equilibrium is reported in the experimental literature.

6 Appendix 1: Existence of reciprocity equilibria

In the presented form, the existence of a reciprocity equilibrium is not always guaranteed. A game where a reciprocity equilibrium may not exist is shown in Figure 11.

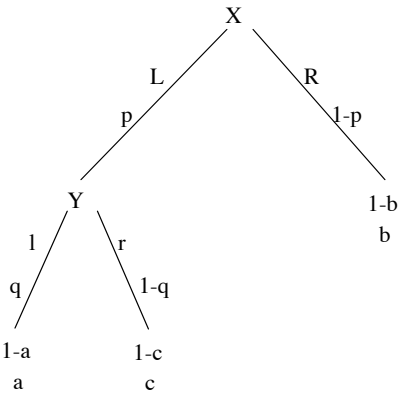

Figure 11: Game tree of a two person game that may not have a reciprocity equilibrium.

Assume that $\frac{1}{2} < a < b < c$ holds. In this game, all possible payoff shares result in an inequity in favor of player 2. Therefore, player 1 is always kind. In node Y player 1's kindness is intentional in this game if and only if $q''a + (1 - q'')c > b$. For instance, if $q'' = 1$, the kindness is not intentional. As a consequence, the reciprocal response will be weak, i.e., the second mover chooses a low q. If, on the other hand, $q'' = 0$, the kindness is intentional and a high q will be chosen. Hence, an equilibrium must lie somewhere in between. However, due to the discontinuity of function Ω, we can find parameters for which there is no reciprocity equilibrium, as claimed in the following proposition.

Proposition 5 *If* $\frac{1}{2b-1} \leq \rho_2 < \frac{1}{(2b-1)\varepsilon_2}$, *there exists no reciprocity equilibrium in the game presented in Figure 11.*

The proof of Proposition 5 is given in Appendix 3.

Figure 12 shows the equilibrium choice of q dependent on ρ_2. As one can see, there is a range (between $1\frac{2}{3}$ and $8\frac{1}{3}$) where there is no equilibrium.

As pointed out above, the reason why the existence of an equilibrium is not guaranteed has to do with the discontinuity of function Ω. To show this, we define for a (small) positive number λ a continuous approximation Ω^λ for Ω.[21] We set

[21] We have $\Omega(\pi_i, \pi_j, \pi_i^0, \pi_j^0) = \lim_{\lambda \to 0} \Omega^\lambda(\pi_i, \pi_j, \pi_i^0, \pi_j^0)$ for any choice of $\pi_i, \pi_j, \pi_i^0, \pi_j^0$.

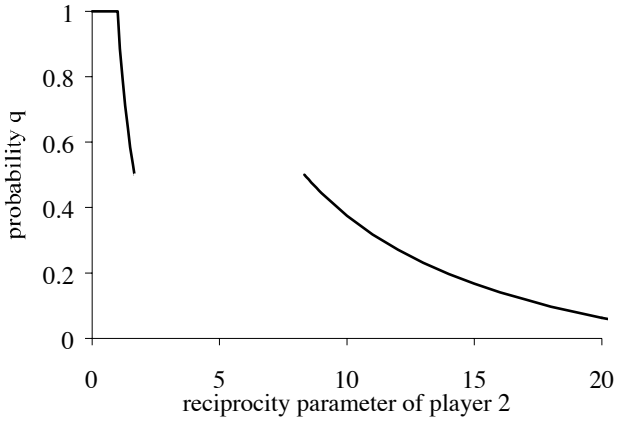

Figure 12: Equilibrium choice of q dependent on ρ_2 for the given parameters $a = 1$, $b = 0.8$, $c = 0.6$, $\varepsilon_2 = 0.2$.

$$\Omega^\lambda(\pi_i, \pi_j, \pi_i^0, \pi_j^0) := \begin{cases} \min(1, \varepsilon_i + \frac{1}{\lambda}(\pi_i^0 - \pi_i)) & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i < \pi_i^0 \\ \varepsilon_i & \text{if } \pi_i^0 \geq \pi_j^0 \text{ and } \pi_i \geq \pi_i^0 \\ \min(1, \varepsilon_i + \frac{1}{\lambda}(\pi_i - \pi_i^0)) & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i \leq \pi_j \\ \max(\varepsilon_i, \min(1 - \frac{\pi_i - \pi_j}{\pi_i^0 - \pi_j^0}, & \text{if } \pi_i^0 < \pi_j^0,\ \pi_i > \pi_i^0 \text{ and } \pi_i > \pi_j \\ \quad \varepsilon_i + \frac{1}{\lambda}(\pi_i - \pi_i^0)) & \\ \varepsilon_i & \text{if } \pi_i^0 < \pi_j^0 \text{ and } \pi_i \leq \pi_i^0 \end{cases}$$

Given the continuous variant Ω^λ of Ω, we define a modified kindness function φ^λ and a utility function U^λ. We call a λ-reciprocity equilibrium a subgame perfect equilibrium of the psychological game with utility U^λ. This modification now guarantees the existence of an equilibrium as the following theorem shows:

Theorem 6 (Existence Theorem) *Let Γ be a finite two person extensive form game with complete information. Let $\lambda > 0$. Then Γ has a λ-reciprocity equilibrium.*

Proof of the Existence Theorem: Let $n \in N_i$ be a node of the game. Then $S^n = \{(p_a)_{a \in A_n} \mid p_a \in [0,1], \sum_a p_a = 1\}$ is the set of mixed strategies in this node. Let $S = \prod_{n \in N} S^n$ be the set of behavior strategy combinations. It includes the strategies of both players. Let $S^{-n} = \prod_{m \neq n} S^m$ be the strategies at all other nodes. Let $s = (s^n, s^{-n})$ be a behavior strategy combination with $s^n \in S^n$ and $s^{-n} \in S^{-n}$. Let s' and s'' the beliefs of first and second order. We define $V(n, (s^n, s^{-n}), s', s'')$ as the utility U_i^λ conditional on node n. If $n \in N_j$ we define $V(n, (s^n, s^{-n}), s', s'')$ as the utility U_j^λ conditional on node n. We now define the best reply correspondence

$$B : S \rightrightarrows S : s \mapsto B(s) \subset S$$

The component $B^n : S \rightrightarrows S^n$ is defined as

$$B^n(s) := \arg\max_{\tau^n \in S^n} V(n, (\tau^n, s^{-n}), s, s)$$

We get

$$B(s) := \{(b^n)_{n\in N} \mid b^n \in B^n(s)\}$$

This definition of the best reply correspondence is related to the agent-strategic form (see FUDENBERG AND TIROLE (1991)): The player behaves as if there was an agent in every decision node. A behavior strategy that optimizes in the agent-strategic form corresponds to a subgame perfect Nash equilibrium.

As S is the product of convex and compact sets, it is convex and compact. Since S is compact and V is continuous in s^n, $B(s)$ is not empty. Because U^λ is linear in the strategies, $B(s)$ is convex.

We now show that B is upper-hemi continuous: First, we show that V depends continuously on the strategies and beliefs: The inequity $\Delta(n)$ is continuous and $\Delta(n) = 0$ holds for $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. Because Ω^λ is bounded (by 1), we get the desired continuity of $\vartheta(n)\Delta(n)$ at $\pi_i(n, s_i'', s_i') = \pi_j(n, s_i'', s_i')$. By construction of Ω^λ, $\vartheta^\lambda(n)$ is continuous in strategies and beliefs if $\pi_i(n, s_i'', s_i') \neq \pi_j(n, s_i'', s_i')$. Therefore, this also holds for $\vartheta^\lambda(n)\Delta(n)$. Hence, $\varphi^\lambda(n)$ is continuous. The reciprocation term and the material profit are obviously continuous and therefore V is continuous. If any function $f : X \times Y \to \mathbb{R}$ is continuous, then the best reply correspondence $R : X \rightrightarrows Y : x \mapsto \arg\max_y f(x, y)$ is upper hemi continuous. Hence, because V is continuous, the best reply correspondence B is upper hemi-continuous.

Therefore, we can apply the fixed point theorem of Kakutani and get an $s^* \in S$ with $s^* \in B(s^*)$. This strategy s^* with first order belief s^* and second order belief s^* now forms a reciprocity equilibrium: The triple (s^*, s^*, s^*) trivially satisfies the consistency of the beliefs. By construction of B each player optimizes his utility in each node - given the beliefs and given the strategies of the other players. This is exactly the definition of the subgame perfect psychological equilibrium. This concludes the proof. ■

7 Appendix 2: Extension to n-person games

The idea behind the generalization to n persons consists in independently considering the reciprocity relation towards each of the other players. We do this in this appendix very briefly.

Let n be the number of players. Let $s_i \in S_i$ be player i's behavior strategy and let $s_i^{(j)} \in S_j$ be player i's first order belief about player j's strategy. Further let $s_i^{(jk)} \in S_k$ be player i's (second order) belief about what he thinks is player j's belief about k's strategy. For $n = 2$, we get $s_i' = s_i^{(j)}$ and $s_i'' = s_i^{(ji)}$. Furthermore, we use the notation $(-ij)$ to express the set of players other than players i and j. So, $s_i^{(j(-ij))}$ is player i's belief about what j believes what the other players will do. In analogy to the definitions (1) to (7) we define - for a given set of first and second order beliefs - the following expressions:

$$\Delta_{j\to i}(n) := \pi_i(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) - \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) \tag{19}$$

$$\Pi_i^j := \left\{ \left(\pi_i(n, s_i^{(ji)}, s_j^p, s_i^{(j(-ij))}), \pi_j(n, s_i^{(ji)}, s_j^p, s_i^{(j(-ij))}) \right) | s_j^p \in S_j^p \right\} \tag{20}$$

$$\vartheta_{j\to i}(n) := \max\left\{\Omega(\pi_i, \pi_j, \pi_i(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}), \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))})) \mid (\pi_i, \pi_j) \in \Pi_i\right\} \tag{21}$$

$$\varphi_{j\to i}(n) = \vartheta_{j\to i}(n)\Delta_{j\to i}(n) \tag{22}$$

$$\sigma_{i\to j}(n, e) := \pi_j(\nu(n, e), s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) - \pi_j(n, s_i^{(ji)}, s_i^{(j)}, s_i^{(j(-ij))}) \tag{23}$$

$$U_i(e) = \pi_i(e) + \rho_i \sum_{j\neq i} \sum_{\substack{n\to e \\ n\in N_i}} \varphi_{j\to i}(n)\sigma_{i\to j}(n, e) \tag{24}$$

A reciprocity equilibrium is again a set of strategies and first and second order believes such that the strategies maximize the utility in Equation (24) and such that strategies and beliefs are consistent.

8 Appendix 3: Proofs of the propositions

Proof of Proposition 1: Let p denote the probability that the responder accepts the offer; p' is the proposer's belief about p and p'' the responder's belief about p'. Let $\vartheta(c)$ be the intention factor at the node after player 1's choice of c. According to Definition 3, the responder's utility in the reciprocity game is given by U_{2A} in case she accepts the offer and by U_{2R} if she rejects the offer:

$$U_{2A} = c + \rho_2\vartheta(c)(p''(c - (1 - c))) \cdot ((1 - c) - p''(1 - c))$$

$$U_{2R} = 0 + \rho_2\vartheta(c)(p''(c - (1 - c))) \cdot (0 - p''(1 - c))$$

The difference in the responder's utility between accepting and rejecting amounts to:

$$U_{2A} - U_{2R} = c + \rho_2\vartheta(c)p''(2c - 1)(1 - c) \tag{25}$$

For $c \geq \frac{1}{2}$, the expression $U_{2A} - U_{2R}$ is positive. Thus, offers of at least half of the pie are always accepted, i.e., $p = 1$. This proves the second part of equation (8) of Proposition 1.

Let us now turn to the case where $c < \frac{1}{2}$. In this case, the second summand in equation (25) is negative. We set equation (25) equal to zero, solve it for p'', and define this solution as p''_{crit}. We get:

$$p''_{crit} = \frac{c}{\rho_2\vartheta(c)(1 - 2c)(1 - c)} \tag{26}$$

If $p'' < p''_{crit}$ then $U_{2A} > U_{2R}$. This implies that player 2 accepts, i.e., $p = 1$. This is an equilibrium if and only if $1 = p = p'' < p''_{crit}$. If, on the other hand, $p'' > p''_{crit}$ then $U_{2A} < U_{2R}$. This implies that player 2 rejects, i.e., $p = 0$. This, however, is impossible because it would imply that $0 = p = p'' > p''_{crit} > 0$. (The last inequality holds because $c < \frac{1}{2}$.) If, finally, $p'' = p''_{crit}$, then $U_{2A} = U_{2R}$. Therefore, in this case,

player 2 is indifferent between accepting and rejecting and we get an equilibrium $p = p'' = p''_{crit}$. This mixed equilibrium is possible for $0 \leq p''_{crit} \leq 1$. Taken together, we have shown that $p^* = \left[\frac{c}{\rho_2 \vartheta(c)(1-2c)(1-c)}\right]^{1]}$.

Let us now discuss the role of intentions. If $c < \frac{1}{2}$ then player 2 is in the disadvantageous situation. Therefore, player 1's move is considered as fully intentional if and only if there is an action of player 1 that leads to a higher payoff for player 2 without bringing player 1 into the disadvantageous situation. A contribution of $c = \frac{1}{2}$ satisfies this condition. Therefore, $\vartheta(c) = 1$ which completes the proof of equation (8).

Equation (8) yields the acceptance probability as a function of c. Let us denote this function as $p(c)$. Furthermore, we define $c_0 := \frac{1+3\rho_2-\sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}$. It is the smallest contribution c where p equals 1. We now turn to the proposer's behavior. His utility according to Definition 3 amounts to:

$$U_1 = p(c) \cdot (1-c) + \rho_1 \vartheta \cdot p(c'') \cdot (1-2c'') \cdot (p(c)c - p(c'')c'') \tag{27}$$

In Equation (27) we applied already the consistency condition for p ($p = p' = p''$) in order to avoid confusion between p' as belief and the derivation of p as a function of c. We first note that Equation (27) is decreasing in c for $c'' \geq \frac{1}{2}$. Therefore, we can conclude that $c \leq \frac{1}{2}$. In this case, player 1 is in the advantageous situation and player 2's move is fully intentional if she has any alternative action that leads to a smaller payoff for player 1. For $c > 0$, this is always the case because player 2 can decrease player 1's payoff with respect to player 1's expectations by rejecting the offer. As we will see below, player 1's contribution is positive if both players have a positive reciprocity parameter. Therefore, we can assume full intentionality, i.e., $\vartheta = 1$.

It can be shown from equation (27) that the proposer's utility is increasing in c as long as his offer is strictly smaller than c_0 (and therefore the acceptance probability smaller than one). Thus, the optimal offer c^* of the responder is at least as high as c_0, i.e., $c^* \geq c_0$.

We now know that the proposer will choose c^* at least as high as to guarantee an acceptance probability of one. Therefore, we consider the situation where the proposer chooses an offer above c_0 (which is always accepted). If $p = 1$, we get:

$$U_{1(p=1)} = (1-c) + \rho_1 \cdot (1-2c'') \cdot (c - c'')$$

$$\frac{\partial U_{1(p=1)}}{\partial c} = -1 + \rho_1 \cdot (1-2c'')$$

We define

$$c''_{crit} = \frac{1}{2}\left(1 - \frac{1}{\rho_1}\right)$$

The utility U_1 is decreasing in c for $c'' > c''_{crit}$, U_1 is increasing in c for $c'' < c''_{crit}$, and U_1 is constant in c for $c'' = c''_{crit}$. Consider the case $c''_{crit} < c_0$: Since in equilibrium $c = c''$, we get $c''_{crit} < c_0 \leq c = c''$. Therefore U_1 is decreasing in c and $c^* = c_0 \, (= \max(c_0, c''_{crit}))$. Consider now the case $c''_{crit} \geq c_0$: If $c'' > c''_{crit}$, U_1 is decreasing in c and therefore c is chosen equal to c_0 which is incompatible with $c = c''$ because $c'' > c''_{crit} \geq c_0 = c$. If $c'' < c''_{crit}$, U_1 is increasing in c and therefore c is chosen

equal to 1 which is also incompatible with $c = c''$ because $c'' < c''_{crit} < \frac{1}{2} < 1 = c$. Therefore $c^* = c'' = c''_{crit} (= \max(c_0, c''_{crit}))$.

This completes the proof of Proposition 1. ■

Proof of Proposition 2:

If $\rho_2 = 0$ the worker chooses $e = 0$ since her material payoff is decreasing in e. Given this, the kindness of player 2 is smaller or equal to zero. Therefore, an increase in w reduces player 1's material payoff and his reciprocity utility. Hence, $w^* = e^* = 0$ is the unique equilibrium for $\rho_2 = 0$.

Let us now assume that $\rho_2 > 0$.

For $w = 0$, the worker chooses an effort of zero. If $w > 0$, the worker always chooses an effort such that her payoff is strictly greater than the payoff of the employer. If the worker chooses $e^* = 0$, then this is obviously the case. If this would not be the case for $e^* > 0$, the worker could increase her utility by reducing her effort. This would increase her own payoff and decreases the payoff of the employer, both leading to an increase in her utility. Consequently, the employer is always kind to the worker and the worker is always unkind to the employer.

The worker's utility is:

$$U_2 = w - \alpha e^2 + \rho_2 \vartheta((w - \alpha e''^2) - (e'' - w))(e - w - (e'' - w))$$

We differentiate U_2 with respect to e, then set $e = e''$ and we get

$$e^* = \left[\frac{-(2\alpha + \rho_2\vartheta) + \sqrt{(2\alpha + \rho_2\vartheta)^2 + 8\alpha(\rho_2\vartheta)^2 w}}{2\alpha\rho_2\vartheta} \right]^{1]}$$

We see that for a positive wage, the worker provides a positive effort. Now, we show that if the employer pays a positive wage, this is fully intentional which implies $\vartheta = 1$. This concludes the proof of formula (11): Indeed, if we take the effort function $e^*(w)$ as given, it can be shown that if $\alpha < \frac{3}{8}$ (which is true since we assumed $\alpha \leq \frac{1}{4}$), the worker's payoff is strictly increasing in the wage w paid. (To see this, calculate $\frac{\partial \pi_2}{\partial w}$.) This fact is independent of the reciprocity parameter ρ_2 and the intention factor ϑ. Therefore, if the employer is kind and pays a wage above zero, the worker gets more than if the employer had paid a wage equal to zero. Therefore, the employer can always be less kind by paying a wage of zero which implies that the employer's kindness is fully intentional.

We now turn to the employer's behavior. As we have seen above, the worker is unkind if the employer pays a positive wage. This unkindness is fully intentional if and only if the effort provided is strictly smaller than 1. (If $e < 1$, the worker could - without switching into a disadvantageous situation - improve the employer's profit by choosing a slightly higher effort. This is impossible if $e = 1$.) Therefore, there are two possible reciprocity equilibria: (i) The employer chooses the minimal wage that guarantees an effort of 1. The worker indeed chooses an effort of 1. In this case, the unkindness of the worker is unintentional. (ii) The employer pays a wage such that the worker provides an effort smaller than 1. In this case the worker actually chooses her effort fully intentionally.

Let us first consider the case in which the employer pays the minimal wage that guarantees an effort of 1. We have defined this wage as $\overline{w}$. To check under which conditions the employer chooses this strategy, we check whether a reduction of the wage would reduce the employer's utility. We calculate

$$\overline{w}(\alpha, \rho_2) = \frac{1}{2\rho_2}\left(2\alpha + \rho_2 + \alpha\rho_2\right)$$

$$\frac{\partial e}{\partial w} = \frac{2\rho_2}{\sqrt{\left(4\alpha^2 + 4\alpha\rho_2 + \rho_2^2 + 8\alpha\rho_2^2 w\right)}}$$

$$\frac{\partial e}{\partial w}(w = \overline{w}) = \frac{2\rho_2}{2\alpha + \rho_2 + 2\alpha\rho_2}$$

and get

$$U_1 = e(w) - w + \rho_1\varepsilon_1(e(w'') + \alpha e(w'') - 2w)(w - \alpha e(w)^2 - (w'' - \alpha e(w'')^2))$$

and

$$\frac{\partial U_1}{\partial w} = \frac{\partial e(w)}{\partial w} - 1 + \rho_1\varepsilon_1(e(w'') + \alpha e(w'') - 2w)(1 - 2\alpha e(w)\frac{\partial e(w)}{\partial w})$$

Now we set $\frac{\partial U_1}{\partial w}(\overline{w}) = 0$ and get a critical

$$(\rho_1\varepsilon_1)_{crit} = \frac{\rho_2\left(-\rho_2 + 2\alpha + 2\alpha\rho_2\right)}{2\alpha\left(-2\alpha - \rho_2 + 2\alpha\rho_2\right)}$$

If the actual value of $\rho_1\varepsilon_1$ is below this critical value, $\overline{w}$ is a local optimum of the employer's utility. This local maximum is global since e is a concave function in w.

This establishes the existence of the first type of equilibria.

It now remains to consider the case where the workers behave fully intentional, i.e., $e(w^*) < 0$ and $\vartheta = 1$. The calculation of this equilibrium leads to $\widetilde{w}(\alpha, \rho_1, \rho_2)$, with

$$\widetilde{w}(\alpha, \rho_1, \rho_2) = \begin{cases} \frac{-4\alpha^2 - 4\alpha\rho_2 + 3\rho_2^2}{8\alpha\rho_2^2} & \text{if } \rho_1 = 0 \\ \frac{3}{16\alpha} + \frac{3}{4\rho_1} + \frac{3\alpha}{4\rho_2^2} + \frac{3}{4\rho_2} + \frac{\rho_2}{8\alpha\rho_1} + \frac{\rho_2^2}{16\alpha\rho_1^2} - & \\ -\frac{(6\alpha\rho_1 + 3\rho_1\rho_2 + \rho_2^2)\sqrt{4\alpha^2\rho_1^2 + 4\alpha\rho_1^2\rho_2 + 12\alpha\rho_1\rho_2^2 + \rho_1^2\rho_2^2 - 2\rho_1\rho_2^3 + \rho_2^4}}{16\alpha\rho_1^2\rho_2^2} & \text{if } \rho_1 > 0 \end{cases}$$

If $\widetilde{w}(\alpha, \rho_1, \rho_2) \leq \overline{w}(\alpha, \rho_2)$, we get $w^* = \widetilde{w}(\alpha, \rho_1, \rho_2)$. If $\widetilde{w}(\alpha, \rho_1, \rho_2) > \overline{w}(\alpha, \rho_2)$ the employer chooses $w^* = \overline{w}(\alpha, \rho_2)$ because for wages above $\overline{w}(\alpha, \rho_2)$ the employer's material payoff as well as the reciprocity utility are decreasing in w. This concludes the proof. ■

Proposition 7 *(Best-shot game) In the normalized reduced best-shot game the reciprocity equilibria are described as follows:*

If $\rho_2\varepsilon_2 < \frac{5}{3}$ then there is a unique reciprocity equilibrium in the reduced best-shot game which is given by

$$r^* = 1$$

$$q^* = \left[\frac{5}{12}\frac{1}{\varepsilon_2\rho_2}\right]^{1]} \tag{28}$$

If $\rho_2\varepsilon_2 < \frac{5}{3}$

$$p^* = \left[\frac{1+\frac{0.8q^*-0.2}{0.6(0.8-0.2q^*)\rho_1}}{1+q^*}\right]^{1]}_{[0} \tag{29}$$

If $\rho_2\varepsilon_2 = \frac{5}{3}$ *and* $\rho_1 = 0$ *then any* $p^* \in [0,1]$ *is part of an equilibrium.*
If $\rho_2\varepsilon_2 = \frac{5}{3}$ *and* $\rho_1 > 0$ *then* $p^* = \frac{4}{5}$.
If $\rho_2\varepsilon_2 > \frac{5}{3}$ *then*

$$p^* = \begin{cases} \left[\frac{1+\frac{0.8q^*-0.2}{0.6(0.8-0.2q^*)\rho_1}}{1+q^*}\right]^{1]}_{[0} & \text{if } \rho_1 \leq \frac{10-40q^*}{12-15q^*+3q^{*2}} \\ \left[\frac{2+\frac{0.8q^*-0.2}{0.6(0.8-0.2q^*)\rho_1}}{3+q^*}\right]^{1]}_{[0} & \text{if } \frac{10-40q^*}{12-15q^*+3q^{*2}} \leq \rho_1 \leq \frac{5(3-11q^*-4q^{*2}+\varepsilon_1(-1+3q^*+4q^{*2}))}{3\varepsilon_1(4-5q^*+q^{*2})} \\ \left[\frac{1+\frac{0.8q^*-0.2}{0.6(0.8-0.2q^*)\varepsilon_1\rho_1}}{1+q^*}\right]^{1]}_{[0} & \text{if } \frac{5(3-11q^*-4q^{*2}+\varepsilon_1(-1+3q^*+4q^{*2}))}{3\varepsilon_1(4-5q^*+q^{*2})} \leq \rho_1 \end{cases} \tag{30}$$

Proof of Proposition 3 and 7 : Let us first consider the situation of player 2 in which she has to decide whether to accept or reject some distribution with a_1 for player 1 and a_2 for player 2. Let U_A denote the utility of player 2 if she accepts and U_R the utility of player 2 if she rejects the offer. Let t be the acceptance probability and n the node where player 2 has to decide. We get

$$U_A = a_2 + \rho_2\vartheta(n)(t''(a_2-a_1))(a_1-\pi_1''(n))$$

$$U_R = 0 + \rho_2\vartheta(n)(t''(a_2-a_1))(0-\pi_1''(n))$$

and

$$U_A - U_R = a_2 + \rho_2\vartheta(n)(t''(a_2-a_1))a_1$$

We see immediately that if $a_2 \geq a_1$ the offer is always accepted. Therefore, $r^* = 1$ in both reduced games.

First, q^*, is derived analogously to the proof of Proposition 1:

$$q^* = \left[\frac{0.2}{0.8\rho_2\vartheta(Y)(0.8-0.2)}\right]^{1]} = \left[\frac{5}{12\rho_2\vartheta(Y)}\right]^{1]}$$

We now have to calculate the value of $\vartheta(Y)$, the intention factor. In the reduced ultimatum game it is equal to 1 because the proposer has the possibility to offer $(0.5, 0.5)$ which is less unkind than $(0.8, 0.2)$ and which is reasonable because the proposer still receives at least as much as the responder. This proves (14).

In the reduced best-shot game the alternative strategy of the proposer $(0.2, 0.8)$ puts himself into a disadvantageous situation. This disadvantageous inequity is at least as large as the advantageous inequity of the reference choice of $(0.8, 0.2)$ (it is

0.6 compared to $0.6q'$). The alternative $(0.2, 0.8)$ has therefore an intention factor of ε_2. Hence, in this case, $\vartheta(Y) = \varepsilon_2$ which proves (17).

In the reduced ultimatum game the second mover is always kind and could be less kind by rejecting the offer. Therefore, player 1's optimal strategy can be calculated with an intention factor of 1. We get

$$U_{82} = 0.8q + \rho_1(0.8p''q + 0.5(1-p'') - (0.2p''q + 0.5(1-p'')))(0.2q - \pi_2'')$$

$$U_{55} = 0.5 + \rho_1(0.8p''q + 0.5(1-p'') - (0.2p''q + 0.5(1-p'')))(0.5 - \pi_2'')$$

$$U_{82} - U_{55} = (0.8q - 0.5) + \rho_1(0.8 - 0.2)p''q(0.2q - 0.5)$$

To get (15), we set $U_{82} - U_{55} = 0$ and apply the same arguments as in the proof of Proposition 1. This proves equation (15).

In the reduced best-shot game, the situation is somewhat more complicated. We get

$$U_{82} = 0.8q + \rho_1\vartheta(X)(0.8p''q + 0.2(1-p'') - (0.2p''q + 0.8(1-p'')))(0.2q - \pi_2'')$$

$$U_{28} = 0.2 + \rho_1\vartheta(X)(0.8p''q + 0.2(1-p'') - (0.2p''q + 0.8(1-p'')))(0.8 - \pi_2'')$$

$$U_{82} - U_{28} = (0.8q - 0.2) + \rho_1\vartheta(X)((0.8 - 0.2)(p''q - (1-p'')))(0.2q - 0.8)$$

and therefore the equilibrium acceptance probability p^* is:

$$p^* = \left[\frac{1 + \frac{0.8q^* - 0.2}{0.6(0.8 - 0.2q^*)\vartheta(X)\rho_1}}{1 + q^*}\right]_{[0}^{1]}$$

For the value of $\vartheta(X)$, we have to consider the intention of player 2. First, we show that in equilibrium $\Delta(X)$ is positive if and only if $0.8q^* - 0.2$ is positive: We get $\Delta(X) = 0.6(p^*q^*) - 0.6(1-p^*)$. If p^* is strictly within the interval $[0,1]$, then $\Delta(X)$ is positive if and only if $0 < p^*q^* - (1-p^*) = p^*(q^*+1) - 1 = \frac{0.8q^* - 0.2}{0.6(0.8 - 0.2q^*)\rho_1}$. Therefore, in this case, $\Delta(X)$ is positive if and only if $0.8q^* - 0.2$ is positive. If $p^* = 1$ then $\Delta(X) = 0.6 * (p^*q^* - (1-p^*)) = 0.6q^*$, which implies $\Delta(X) > 0$. Note that $p^* = 1$ can only occur if $0.8q^* - 0.2 > 0$. If $p^* = 0$ then $\Delta(X) = 0.6 * (p^*q^* - (1-p^*)) = -0.6$, therefore $\Delta(X) < 0$. This can only be the case if $0.8q^* - 0.2 < 0$.

If $0.8q^* - 0.2 > 0$, i.e., $q^* > \frac{1}{4}$, player 2 is kind and can be less kind by rejecting any offer. Therefore, in this case, $\vartheta(X) = 1$. This proves (29).

If $q^* = \frac{1}{4}$, then player 1 is materially indifferent between his two possible actions. If $\rho_1 = 0$, he is also indifferent with respect to his reciprocity utility. If $\rho_1 > 0$, then only $p^* = \frac{4}{5}$ can be part of an equilibrium: This choice leads to an equitable expected payoff of $(0.2, 0.2)$. Any other choice of p^* would lead to some inequity. Because player 1 is materially indifferent between the two choices, he would reciprocate such inequity and try to produce an inequity in the opposite direction. Therefore, $p^* = \frac{4}{5}$ is the only equilibrium choice in this case.

If $0.8q^* - 0.2 < 0$, i.e., $q^* < \frac{1}{4}$, player 2 is unkind but could be less unkind if she would accept both offers, i.e., if she would choose $q = 1$. This is reasonable if $p \leq \frac{1}{2}$

because in this case, player 2 is still in the advantageous position. The inequality $p \leq \frac{1}{2}$ is satisfied if $\rho_1 \leq \frac{10-40q^*}{12-15q^*+3q^{*2}}$. If $p > \frac{1}{2}$, the intention factor is

$$\vartheta(X) = \max\left(\varepsilon_1, 1 - \frac{-0.6p'' + 0.6(1-p'')}{0.6p''q' - 0.6(1-p'')}\right) \tag{31}$$

If the first term applies, we get

$$p^* = \left[\frac{1 + \frac{0.8q^*-0.2}{0.6(0.8-0.2q^*)\varepsilon_1\rho_1}}{1+q^*}\right]_{[0}^{1]}$$

The first term applies if $1 - \frac{-0.6p^*+0.6(1-p^*)}{0.6p^*q^*-0.6(1-p^*)} \leq \varepsilon_1$ which is the case if $\rho_1 \geq \frac{5(3-11q^*-4q^{*2}+\varepsilon_1(-1+3q^*+4q^{*2}))}{3\varepsilon_1(4-5q^*+q^{*2})}$. Finally, if $\rho_1 < \frac{5(3-11q^*-4q^{*2}+\varepsilon_1(-1+3q^*+4q^{*2}))}{3\varepsilon_1(4-5q^*+q^{*2})}$ then in (31), the second term applies and

$$p^* = \left[\frac{2 - \frac{(0.2-.8q^*)}{0.6(0.8-0.2q^*)\rho_1}}{3+q^*}\right]_{[0}^{1]}$$

which concludes the proof. ■

Proof of Proposition 4:

First note that the responder will accept any offer above 0.5. This holds because accepting yields a higher material payoff than rejecting. Moreover, accepting provides a higher reciprocity utility with respect to the proposer who offers $c_{\max}$ because this proposer is kind towards the responder. Furthermore, the responder's reciprocity utility with respect to the other proposers is zero, independent of the responder's decision.

If at least two players offer $c = 1$ then for all proposers any offer is a best response since all payoffs are unaffected by their decisions. Therefore, the strategies described in Proposition 4 are indeed an equilibrium. Only one proposer offering $c = 1$ cannot be an equilibrium since this player has an incentive to lower his offer slightly: This increases his material payoff and the reciprocity utility towards the responder. Changing the offer from $c = 1$ has no effect on the reciprocity utility towards the other responders because $\Delta = 0$. If $c_{\max} < 1$ then all proposers have an incentive to increase their offer to $c_{\max}$. If at least two proposers offer $c_{\max} < 1$ then these proposer have an incentive to increase their offer infinitesimally. This increases the material payoff from $\frac{c_{\max}}{k}$ to $c_{\max}$, where k is the number of proposers offering $c_{\max}$. On the other hand, it changes the reciprocity utilities only infinitesimally. Hence $c_{\max} < 1$ cannot be an equilibrium. ■

Proof of Proposition 5:

To proof Proposition 5 we show that there is no q that can be part of an equilibrium:

The utility of player 2 in the end nodes Ll and Lr is:

$$\begin{aligned} U_{2Ll} &= a + \rho_2\vartheta(q''(a-(1-a)) + (1-q'')(c-(1-c))(1-a) \\ U_{2Lr} &= a + \rho_2\vartheta(q''(a-(1-a)) + (1-q'')(c-(1-c))(1-c) \end{aligned}$$

We set:

$$\begin{aligned} U_{2Ll} - U_{2Lr} &= (c-a) + \rho_2 \vartheta (q''(2a-2c) + (2c-1)(a-c)) \\ &= (c-a)\left[1 - \rho_2 \vartheta (2c - 1 - q''(2a-2c))\right] \end{aligned}$$

We define:

$$q''_{crit} = \frac{c - \frac{1}{2}}{c-a} - \frac{1}{\rho_2 \vartheta 2(c-a)}$$

This is the solution of the equation $U_{2Ll} - U_{2Lr} = 0$. Player 2 will choose $q = 0$ if $q'' > q''_{crit}$ and $q = 1$ if $q'' < q''_{crit}$. We show that this is impossible for the parameters given in the proposition.

Case 1: If $q''a + (1-q'')c > b$ holds, there is full intention, i.e., $\vartheta = 1$. The inequality $q''a + (1-q'')c > b$ is equivalent to $q'' < \frac{c-b}{c-a}$. Furthermore, $\frac{1}{2b-1} < \rho_2$ holds by assumptions. Hence, we get $q_{crit} \geq \frac{c-\frac{1}{2}}{c-a} - \frac{1}{\frac{1}{2b-1}2(c-a)} = \frac{c-\frac{1}{2}}{c-a} - \frac{b-\frac{1}{2}}{c-a} = \frac{c-b}{c-a} > q''$.

As we have seen above this can only be the case if $q = q'' = 1$ and therefore, $q_{crit} > 1$. But by definition of a, b and c the inequality $\frac{c-b}{c-a} < 1$ holds.

Case 2: If $q''a + (1-q'')c \leq b$ holds, no intentions are involved and $\vartheta = \varepsilon_2$ holds. The inequality $q''a + (1-q'')c \leq b$ is equivalent to $q'' \geq \frac{c-b}{c-a}$. Furthermore, $\frac{1}{(2b-1)\varepsilon_2} > \rho_2$ holds by assumptions and we get $q_{crit} < \frac{c-\frac{1}{2}}{c-a} - \frac{1}{\frac{1}{2b-1}2\varepsilon_2(c-a)} = \frac{c-b}{c-a} \leq \frac{c-b}{c-a} \leq q''$. As we have seen above, this can only be the case if $q = q'' = 0$ and therefore, $q_{crit} \leq 0$. But by definition of a, b and c the inequality $\frac{c-b}{c-a} > 0$ holds. ■

9 References

Adams, J. S. (1965): "Inequity in Social Exchange", in: Leonhard Berkowitz (ed.), Advances in Experimental Psychology 2, New York: Academic Press, 267-299.

Agell, J. (1999): "On the Benefits from Rigid Labour Markets: Norms, Market Failures, and Social Insurance", Economic Journal 109, 143-164.

Agell, J. and Lundborg, P. (1995): "Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms", Scandinavian Journal of Economics 97, 295 - 307.

Berg, J., Dickhaut J. and McCabe K. (1995): "Trust, Reciprocity, and Social History", Games and Economic Behavior 10, 122-142.

Bewley, T. (1995): "A Depressed Labor Market as Explained by Participants", American Economic Review 85, Papers and Proceedings, 250 - 254.

Blinder, A. S. and Choi, D. H. (1990): "A Shred of Evidence on Theories of Wage Stickiness", Quarterly Journal of Economics 105, 1003-1016.

Blount, S. (1995): "When Social Outcomes aren't Fair: The Effect of Causal Attributions on Preferences", Organizational Behavior & Human Decision Processes 63, 131-144.

Bolle, F. and Ockenfels, P. (1990): "Prisoner's Dilemma as a Game with Incomplete Information", Journal of Economic Psychology, 11, 69-84.

Bolle, F. and Kritikos, A. (1998): "Self-Centered Inequality Aversion versus Reciprocity and Altruism" mimeo, 14/95, Europe-University Viadrina, Frankfurt/Oder.

Bolton, G. and Ockenfels, A. (2000): "ERC - A Theory of Equity, Reciprocity and Competition", American Economic Review 90, 166-193.

Camerer, C. and Thaler, R. (1995): "Ultimatums, Dictators, and Manners", Journal of Economic Perspectives 9, 209-219.

Campbell, C. M. and Kamlani, K. (1997): "The Reasons for Wage Rigidity: Evidence from a Survey of Firms", Quarterly Journal of Economics 112, 759-789.

Charness, G. (1996): "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation", mimeo, University of Berkeley.

Clark, K. and Sefton, M. (1998): "The Sequential Prisoner's Dilemma: Evidence on Reciprocal Altruism", mimeo, University of Manchaster.

Davis, D. and Holt, C. (1993): Experimental Economics, Princeton University Press, Princeton.

Dufwenberg, M. and Kirchsteiger, G. (1998): "A Theory of Sequential Reciprocity", mimeo, CentER for Economic Research, Tilburg.

Eckel, C. and Grossman, P. (1998): "Are Women Less Selfish Than Men? Evidence from Dictator Experiments", Economic Journal 108, 726-35.

Falk, A, Fehr E., and Fischbacher U. (1999): "On the Nature of Fair Behavior – Intentions Matter", mimeo, University of Zurich.

Falk, A, Fehr E., and Fischbacher U. (2000): "Informal Sactions – How Much and Why?", mimeo, University of Zurich.

Fehr, E. and Falk, A. (1999): "Wage Rigidities in a Competitive, Incomplete Contract Market", Journal of Political Economy 107, 106-134.

Fehr, E., Gächter, S., and Kirchsteiger G. (1997): "Reciprocity as a Contract Enforcement Device: Experimental Evidence", Econometrica 65, 833-860.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation", Quarterly Journal of Economics 108, 437-460.

Fehr, E. and Schmidt, K. (1999): "A Theory of Fairness, Competition, and Cooperation", Quarterly Journal of Economics 114, August 1999, 817-868.

Fischbacher, U., Gächter, S. and Fehr, E. (1999): "Anomalous Behavior in Public Goods Experiments: How Much and Why: Comment", mimeo, University of Zurich.

Forsythe, R., Horowitz, J., Savin, N. and Sefton, M. (1994): "Fairness in Simple Bargaining Experiments", Games and Economic Behavior 6, 347-369.

Francis, H. (1985): "The Law, Oral Tradition and the Mining Community", Journal of Law and Society 12, 2267-2271.

Fudenberg, D. and Tirole, J. (1991): Game Theory. The MIT Press, Cambridge, Massachussets.

Gächter, S. and Falk, A. (1997): "Reputation or Reciprocity?", mimeo, University of Zurich.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989): "Psychological Games and Sequential Rationality", Games and Economic Behavior 1, 60 - 79.

Gouldner, A. (1960): "The Norm of Reciprocity", American Sociological Review 25, 161 - 178.

Goranson, R. E. and Berkowitz, L. (1966): "Reciprocity and Responsibility Reactions to Prior Help", Journal of Personality and Social Psychology 3, 227-232.

Greenberg, M. S. and Frisch, D. M. (1972): "Effect of Intentionality on Willingness to Reciprocate a Favord", Journal of Experimental Social Psychology 8, 99-111.

Güth, W., Schmittberger, R. and Schwarze, B. (1982): "An Experimental Analysis of Ultimatium Bargaining", Journal of Economic Behavior and Organization 3, 367-88.

Güth, W. (1995): "On Ultimatum Bargaining Experiments - A Personal Review", Journal of Economic Behavior and Organization 27, 329 - 344.

Güth, W., Huck, S. and Ockenfels P. (1996): "Two-level ultimatum bargaining with incomplete information" Economic Journal 106, 593-604.

Güth, W., Marchand, N. and Rullière, J.L. (1997): "On the Reliability of Reciprocal Fairness - An Experimental Study", Discussion paper, Humboldt-University Berlin.

Harrison, G. W. and Hirshleifer J. (1989): "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods", Journal of Political Economy 97, 201-225.

Hayashi, N., Ostrom, E., Walker, J., and Yamagishi, T. (1998): "Reciprocity, Trust, and the Sense of Control: A Cross-Societal Study", Discussion paper, Indiana University, Bloomington.

Hoffman, E., McCabe, K. and Smith, V. L. (1996): "Social Distance and Other-Regarding Behavior in Dictator Games", American Economic Review 86, 653-660.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994): "Preferences, Property Rights, and Anonymity in Bargaining Games", Games and Economic Behavior 7, 346-380.

Kahneman, D., Knetsch, J. and Thaler, R. (1986): "Fairness as a Constraint on Profit-Seeking: Entitlements in the Market", American Economic Review 76, 4, 728-741.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, R. (1982): "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", Journal of Economic Theory 27, 245-252.

Ledyard, J. (1995): "Public Goods: A Survey of Experimental Research", in: Handbook of Experimental Economics, ed. by J. Kagel and A. Roth, Princeton: Princeton University Press.

Levine, D. (1998): "Modeling Altruism and Spitefulness in Experiments", Review of Economic Dynamics 1, 593-622.

Loewenstein, G. F., Thompson, L., and Bazerman, M. H. (1989): "Social Utility and Decision Making in Interpersonal Contexts", Journal of Personality and Social Psychology 57, 426-441.

Miller, S. (1997): "Strategieuntersuchung zum Investitionsspiel von Berg, Dickhaut und McCabe", Diploma thesis, University of Bonn.

Mowday, R. T. (1991): "Equity Theory Predictions of Behavior in Organizations", in: R. M. Steers and L. W. Porter (eds.), Motivation and Work Behavior, New York: McGraw-Hill, 111-130.

Prasnikar, V. and Roth, A. E. (1992), "Considerations of Fairness and Strategy: Experimental Data from Sequential Games", Quarterly Journal of Economics, 865-888.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics", American Economic Review 83, 1281 - 1302.

Rabin, M. (1998): "Psychology and Economics", Journal of Economic Literature 36, 11 - 46.

Roth, A. (1995): "Bargaining Experiments", in: Handbook of Experimental Economics, ed. by J. Kagel and A. Roth, Princeton: Princeton University Press.

Roth, A., Prasnikar, V., Okuno-Fujiwara, M. and Zamir, S. (1991): "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study", American Economic Review 81, 1068-1095.

Ruffle, B. (1995): "Gift Giving with Emotions", mimeo, Dept. of Economics, Princeton University.

Selten, R. (1978): "The Equity Principle in Economic Behavior", In: Decision Theory and Social Ethics. Issues in Social Choice, H. Gottinger and W. Leinfellner (Eds.), Reidel Dordrecht.

Seneca, L.A. (1917): "On Benefits" [De Beneficiis], in: Seneca's Morals, New York, Harper and Row.

Smith, A. (1976): The Theory of Moral Sentiments. Edited by D.D. Raphael and A.C. Macfie, Oxford, Clarendon Press.

Smith, K. W. (1992): "Reciprocity and Fairness: Positive Incentives for Tax Compliance", in: Why People Pay Taxes - Tax Compliance and Enforcement, ed. by Joel Slemrod, The University of Michigan Press, Ann Arbor, 223-250.

Smith, V. L. (1982); "Microeconomic Systems as an Experimental Science", American Economic Review 72, 923-955.

Steers, R. M. and Porter, L. W. (1991): Motivation and Work Behavior, Fifth Edition, New York: McGraw-Hill.

Sugden R. (1984): "Reciprocity: The Supply of Public Goods Through Voluntary Contributions", The Economic Journal 94, 772-787

Thaler, R. H. (1988): "Anomalies: The Ultimatum Game", The Journal of Economic Perspectives 2, 195-206.

Trivers, R. (1971): "The Evolution of Reciprocal Altruism", Quaterly Review of Biology 46, 35-57.

Walster, E. and Walster, G. W. (1978): Equity - Theory and Research, Boston, Allyn and Bacon.

Watabe, M., Terai, S., Hayashi, N., and Yamagishi, T. (1996): "Cooperation in the One-Shot Prisoner's Dilemma Based on Expectations of Reciprocity", Japanese Journal of Experimental Social Psychology, XXXVI, 183-196.